A thesis
for the degree of Master of Science

Algorithm to detect the conservative proteins in species of high mutation rate

높은 변이율을 갖는 종에서의 보존적인 단백질을 감지하는 연산 (algorithm)

August, 2010

By
Haitham Sobhy Abdel-Mohaimen Hassan

School of Agricultural Biotechnology
Graduate School, Seoul National University
농학 석사 학위 논문

높은 변이율을 갖는 종에서의 보존적인 단백질을 감지하는 연산 (algorithm)

Algorithm to detect the conservative proteins in species of high mutation rate

지도 교수　김 희 발

이 논문을 농학 석사 학위 논문으로 제출함

2010 년　8 월
서 울 대 학 교　대 학 원
농 생 명 공 학　전공

하이삼 수비히
Haitham Sobhy

Haitham 의 농학 석사 학위 논문을 인준함.

2010 년　8 월

위 원 장________________________________(인)

부위원장________________________________(인)

위　　원________________________________(인)

A thesis for the degree of Master of Science

Algorithm to detect the conservative proteins in species of high mutation rate

Under the supervision of Professor
Hee-Bal Kim

By
Haitham Sobhy Abdel-Mohaimen Hassan

A thesis submitted to Graduate School
Department of Agricultural Biotechnology
Seoul National University
In partial fulfillment of the requirements for
the degree of Master of Science

August, 2010

ACKNOWLEDGEMENT

I would like to extend my immense appreciation and my heartfelt thanks to my supervisor, Dr. **Hee-Bal Kim** for his support, advices, comments, and guidance that encouraged me during my research work.

I heartily thank Dr. **Cheol-Heui Yun** for his great advices, and the valuable guidance. I am very grateful for his support, the great discussions, and the positive advices.

Moreover, I like to thank the third member of evaluation committee, Dr. **Jong-Sik Chun** for his great comments, and guidance.

My special thanks to Mr. **Jong-Eun Park** for his great ideas, and strong co-operation that aid to accomplish this study.

The great comments, and support from Dr. **Bo-Young Lee**, Dr. **Ehab Mostafa**, Mrs. **Sam-Sun Sung**, Mrs. **Santi Upadhaya**, and Dr. **Sook-Hee Yoon** helped me to improve my research.

I would like to thank my laboratory members, who have helped me during my staying in Korea.

I like to express my special respect to the **National Institute for International Education** (**NIIED**) for providing fund to study my master degree in Korea.

The most importantly, I would like to offer my special thanks from my heart to **my family members** for their forbearance, generous support, and encouraging me during my whole life. Without all of which it would not have been possible to achieve my goals.

SUMMARY

Finding a phylogenetic marker in species plays the important role to uncover the evolutionary relationship of the species. There are numbers of methods for whole-genome comparisons require with many limitations; moreover, bacteriophage genomes cannot be easily analyzed by these methods. This work suggests better approach and aims to provide an automated prediction system of phylogenetic marker.

A new application called "CoMark" has been developed to theoretically predict the conservative proteins over the completed genomes. These predicted proteins are applicant to be phylogenetic marker for the strains that used by the program.

The prediction system was initialized by searching a large collection of whole protein sequences into groups based on their protein sequence similarity. From each individual group, each data set was then aligned and used to determine an evolutionary distance between all pairs of closely phage proteins that calculated the average similarity between genomes. Then generated phage proteome tree was compared with all tree s which are constructed using each data set. Experiments were conducted with 5797 (5821 including query) protein sequences over 96 (97 including query) bacteriophage genomes.

The proposed system employed proteomic context to predict a genetic marker and could detect marker correspondence in whole-genome comparisons. Although the experiment focused on the strains of bacteriophages, the method might be extended to other microbial

genomes as they shared a number of similar characteristics with phage genomes such as functional protein conservation.

The software is freeware, desktop application for windows system and available with full documentation and user guide at http://snugenome.snu.ac.kr/comark/.

Keywords: Bacteriophage, conservative proteins, horizontal gene transfer, phylogenetic marker, proteins, tree comparison, whole-proteome tree.

Student Number: 2008-22515

CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEWS

1. PHYLOGENETIC MAKER

The Phylogenetic Marker [PM] is a fragment of either coding or non-coding DNA or protein which is used in phylogenetic tree constructions and evolutionary studies of different species.

In the age of metagenomics, and due to the fast increasing molecular sequence data, the importance of studying and classifying the new and unknown species has been increased. Currently, there is an increasing concern to construct the evolutionary tree of different organisms. These will lead to new advances in the field of molecular evolution will be achieved.

However, phylogenetic analysis using sequence data is found to have a limitation to resolve in particular long branches. Furthermore, there are number of basic and main factors which are affecting the branching patterns of any species while constructing the phylogenetic trees, like long-branch attraction effect. These factors may cause a contradictory result or the tree may fail to resolve certain evolutionary relationships between the interested species (Bacteriodetes and Spirochaetes; Gupta, 2005).

Here, one can find the importance of the 'marker genes', or 'phylogenetic markers'. These markers are not only novel motifs but also more reliable types of motifs. These genetic markers are capable of preserving enough information about the evolutionary relationships of these species (Krauss, et al., 2008).

In addition, these markers should have no or small predictable variation within a given species and their sequences are available for most

or all species of a genus. This means that the phylogenetic marker is usually a conservative fragment; this is the most characteristics feature of phylogenetic markers. It important to note that, not all conservative fragments can work as phylogenetic marker.

Moreover, each marker should be distributed in wide range along the genome of the organism and derived from the wealth of genomic data. Furthermore, the length of the gene sequence should be enlarged enough to maintain and include adequate information about the history and taxonomy between different species to track the evolutionary relationships between them (Thompson, et al., 2005; Zeigler, 2003).

The ribosomal RNAs (rRNAs) which are highly conserved genes, this allow them to be one of the most important markers that are widely used in the identification of the source organism (McHardy, et al., 2007). The rRNAs are highly conserved. The phylogenetic marker is used to expose the evolutionary relationships of the species and to classify metagenomic fragments according to taxonomic characterization.

2. BACTERIOPHAGES

Viruses are the very small organism, have genome of small size and at the same time are obligate intracellular parasite that can infect prokaryotes like Bacteria. Therefore, the viruses that infect the bacterial cells are known as Bacteriophages. The viruses that infect another viruses have been also reported and it is known as “Sputnik virophage” (La Scola, et al., 2008).

Bacteriophages, or phages for short, were first described in the early by Frederick Twort in 1915 and Felix d'Herelle in 1917, (Nelson, 2004; Twort, 1915). Phages are most abundant in the biosphere and extremely common in the environment. Up till now, very little information has been revealed about the phages, like their biodiversity, biogeography, or phylogeny. The phages genome characterized by high rate of insertion and deletion (indel) process, and high mutation rate.

The studies carried on the phage model systems played an important role to develop the field of biology and were the basic establishment of the field of molecular biology field; these advances revolutionize the biomedical studies as well. The phage therapy is one of these new approaches. On the other hand, number of diagnostic and therapeutic tools in the drug discovery and development are depending on the phages and their proteome (Nelson, 2004).

Recently, a new studies carried on the influences of phage on ecosystems (Fuhrman, 1999; Fuller, et al., 1998). In addition, the bacteriophages may infect the milk fermenting bacteria like several species of *Lactococcus* or *Lactobacillus* (Proux, et al., 2002). In turn, this may cause serious problem for the milk industry.

The taxonomy of the bacteriophages has been gone through several steps, and several attempts have been done during the past years. One of the most important attempts took place in 1966. In this year, the International Congress of Microbiology launched the International Committee on the Taxonomy of Viruses (ICTV) to establish a universal taxonomy for the viruses (Nelson, 2004).

Analyses of the 16 S ribosomal genes in bacteria, and ribosomal RNA (rRNA) sequences and sequencing of uncultured 16S rRNAs has great

role to revolutionize the field of taxonomy (Fox, et al., 1980; Lane, et al., 1985; Nelson, 2004; Pace, et al., 1986).

As all viruses, and unlike to some bacteria or some other organisms, phages have neither single sequence nor universal ribosomal DNA (rDNA) sequences, or even conserved sequences (genes or proteins) that can detect phylogenetic and taxonomic relationships from them (Nelson, 2004; Rohwer and Edwards, 2002). As consequences of the complicated genome of the phages, it was hard to develop sequence-based taxonomic systems for the phages, this is the reason that there is a huge number of unclassified genomes belong to phage species (Rohwer and Edwards, 2002).

As mentioned above, the bacteriophage genome is characterized by the absence of the common, definite or well known phylogenetic marker such as 16S rRNA to study the phylogeny of phage, and rarely tried to construct to phylogeny for sub-lineages by single marker (Filée, et al., 2005; Fouts, 2006; Tetart, et al., 2001). In addition, other studies claim that the phage genome is characterized by mosaic structure of genome (Fouts, 2006; Lawrence, et al., 2002). As a result of these genetic challenges, the bacteriophage genomes can be classified as complicated genomes, and they are not simply to be analyzed.

Unfortunately, the classification of phages based on genetic markers or sequences neither met with a great success nor advances. The current taxonomic and phylogenetic system developed for phages by ICTV, for example, is based on the physical characteristics of the virion (Nelson, 2004; Rohwer and Edwards, 2002). Several other studies classify the phage based on the structural proteins such as capsids (Fuhrman, 1999; Hambly, et al., 2001; Le Marrec, et al., 1997; SUSSKIND and BOTSTEIN, 1978).

These genetic difficulties may be as consequences of the high diversified nature of phages, and high rate of genetic alteration processes. These factors made the phages have no conserved regions or genetic markers to ease the classification or to maintain the evolutionary relationships.

Therefore, once studies succeed to find a conservative protein or a common genetic marker of phages, this will improve the quality of evolutionary phylogenetic studies of the phages.

3. WHOLE-PROTEOME TREE

Due to the complex structure of the bacteriophage genome, and absence of the common phylogenetic tree of its species, the ICTV c is depend on the physical characters of the phages in a classification of them (Rohwer and Edwards, 2002). This means that there is no sequence based taxonomy of the phages.

Rohwer and colleague (Rohwer and Edwards, 2002) developed a new method to construct the phylogenetic tree of the phages based on the whole-proteome come from proteins of different phages strains. They claim that is a basis of the taxonomical system for phages based on their genomes. They reported that the method is compatible with the roles of the ICTV organization.

Chapter 2 contains a description of the tree construction algorithm, and the figure 2-1 in chapter 2 is describing the main steps of the whole-proteome tree construction.

4. GENE TRANSFER

The gene transfer process is the insertion of the genetic information or nucleic acid, DNA for example, to new a genome of new cell. The nucleic acids or DNA may be transferred to the new cell, or in some cases the naked DNA transfers from one genome to another genome of another cell with already existing genome. It is very important to note that the transfer of genes from an organelle to the nucleus within the same cell has been reported by previous studies (Andersson, 2005).

The gene transfer is mainly divided into two types, depending on the donor and the receptor cells relationships.

4.1. VERTICAL GENE TRANSFER (VGT)

The Vertical Gene Transfer (VGT for short) process is the transfer of genes from the mother cells to daughter cells. The VGT process is responsible for moving the genetic material from ancestors to the offspring; in other words, it is the inheritance of genes from one generation to another. In higher organisms VGT is known as a sexual reproduction or crossing.

Figure 1-1 is the schematic diagram to show the different types of mechanisms of the gene transfer process.

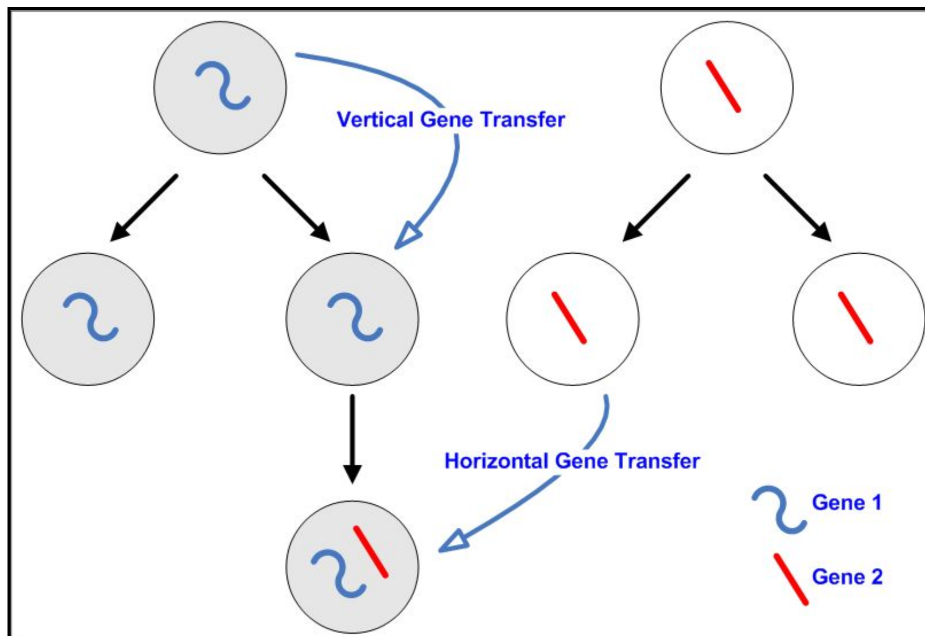


Figure 1-1. A schematic diagram to show the different types of mechanisms of the gene transfer process. Blue curved line is the gene 1 moved vertically from the ancestor to the offspring. The red straight line is representing the gene 2 which is transferred vertically from the ancestor to the offspring and horizontally from one species to another.

4.2. HORIZONTAL GENE TRANSFER (HGT)

At the first, it has been reported that there is a homology between certain genes in one species and another one in another evolutionary distant species. This is phenomenon is attributed to the gene transfer between two different cells from two different species.

Thus, horizontal gene transfer (HGT) is another type of gene transfer process has been discovered that is responsible for the transfer of the genetic materials devoid of any sexual reproduction and without any relation of gene inheritance from the ancestors. The Horizontal Gene Transfer process is also known as Lateral Gene Transfer (LGT) (Brown, 2003; Doolittle, 1999; Kurland, et al., 2003).

The phenomenon is very common in prokaryotes. The most proposed methods of transferring the genes are Transformation, Transduction, or Bacterial conjugation between bacteria and bacteria or other organisms (Doolittle, 1999; Jiang and Paul, 1998; Kidambi, et al., 1994; Lorenz and Wackernagel, 1994).

Transformation is a process of introducing new DNA molecules to the genetic material of the cell. An example is the plasmid transformation in which the uptake a plasmid into the bacteriophage DNA. The phenomenon is common in bacteria and one of the causes of the genetic alteration.

Transduction is moving of the genetic material from bacterial cell, for example, to another through a third party organism usually caused by viruses.

Bacterial conjugation is the transfer of the genetic molecules from one bacterium to another due to the cell-to-cell contact.

Here, due to the high rate of the transformation, transduction, or bacterial conjugation in Archaea and bacteria; these species are highly enriched by the horizontal gene transfer.

Moreover, in 2008, a new virus called "Sputnik" as a virophage (La Scola, et al., 2008). It is also reported the occurrence of the gene transfer process. Other reports claim that the genes could be transferred from one species to another in eukaryotes like fungi or plants by the same mechanism. By the HGT mechanism, the antibiotic resistance genes can be transferred from bacteria into a transgenic plants and then to new bacterial cell (Smalla, et al., 2000).

The transformation of genes may be from species to another species of the distant genome, it means that the two species may be ortholog or homolog. Therefore, a homologous recombination may be carried on between the two copies; then chimerical gene may be resulted (Philippe and Douady, 2003). The chimeric proteins or the fusion proteins are proteins that come from joining of more than one gene, proteins or peptides together.

Here, the genomes of species which composed of genes from multiple sources fail to construct neither universal nor unique phylogenetic tree. In the other word, the HGT mechanism is claimed by some reports to be the answer the anomalies in phylogeny and incongruence of trees in prokaryotes (Brown, 2003; Creevey, et al., 2004; Eisen, 2000; Garcia-Vallve, et al., 2003; Koonin, et al., 2001; Philippe and Douady, 2003; Puigbo, et al., 2007; Ragan, 2001).

Figure 1-1 is the schematic diagram to show the different types of mechanisms of the gene transfer process.

DETECTION AND QUANTIFICATION OF THE HGT PROCESS

Due to the increasing and high importance of HGT mechanism, numbers of methods have been developed to detect it; however, the bioinformatics tool cannot easily detect the HGT process (Brown, 2003; Fall, et al., 2007; Philippe and Douady, 2003):

1. The incongruence of tree is one of the important methods of detection.

2. The other method is deviation of nucleotide composition between neighboring sequences (for example elevation of G+C bases).

3. The unusual or anomalous phylogenetic distribution of one or more genes in one genome, at the same time this genes present in another distant genome but absent of closely related genome is another method for detecting occurrence of HGT.

4. The homology of genes (sequence similarity between two distant genomes) is another category of the detection techniques.

In the same context, there are numbers of new procedures developed to quantify the HGT (Mirkin, et al., 2003; Philippe and Douady, 2003).

4.3. IMPORTANCES OF THE GENE TRANSFER PROCESS

As the HGT process can be done either naturally or artificially in laboratories. The importance of HGT is come from the fact that the mechanism is capable of causing the alteration of the genetic materials in the organisms, especially in the prokaryotes. For example, drug resistant strains can be created by acquiring an antibiotic resistance gene into their genomes. Moreover, not only the evolutionary studies but also the new biotechnological techniques like the plant breeding, and transgenic plants are important implications of the HGT mechanism.

There are several applications of the gene transfer especially the horizontal gene transfer process can be listed as following (Davids and Zhang, 2008; Ochman, et al., 2000):

1. The HGT may play a role in the genomic alteration and then, it will change the evolutionary phylogenetic relationships. The tree incongruence is one of the consequences of the genetic alterations.

2. In most cases, coding or non-coding regions are subjecting to the HGT process. Therefore, it is expected to proteomic products as well subjected to be changed.

3. Since, the proteins products may be changed by HGT, it has been reported that the HGT process may affect the protein-protein interactions and network.

4. In case of occurrence of the HGT process naturally, one gene can be transferred from one bacterial cell to another. The negative impact will appear in the case of the gene is contribute in the pathogenicity of the

microbe. In some other cases the gene might increase the virulence of the microbe.

5. One of the most important implications in the biomedical field is the transfer of the antimicrobial (antibiotic) resistance genes between species.

6. The positive impacts of the natural HGT process will be clear in the case of the transferred gene improve the fitness of the certain population, or increase the desirable properties in case of industrial bacteria. In the same context, the other positive impacts will be achieved if the gene reduces the pathogenicity of the pathogen.

7. One of the most economic implications of HGT is the industrial approaches, like biotechnological and diary industries that depend on the bacteria or protein products derived from bacteria.

8. In the laboratories, the economic impacts these processes are associated with the biotechnological industry. In this case, the desirable treats or genes can be intentionally transferred to the other organism.

9. The plant breeding, genetically modified plants (GMP) and biotechnological crops are one of applications of the artificial HGT.

10. The production of the transgenic animals is also produced by introducing one or more DNA into embryonic stem cells.

11. One of the most famous and important applications of the introducing of the gene of interest into a new genome is the vaccine development, and the drug discovery.

5. TREE COMPARISON

The main goal of the phylogenetic and evolutionary studies is to explain the history of the relationships, similarities and differences between different species. In most cases, the prokaryotes yield phylogenetic trees with abnormalities and incongruent topology. As a consequence of the high mutation, genomic alteration rates on these species genomes. In other words, the construction of accurate phylogenetic tree will help to classify species and reflect their evolutionary historical background.

Therefore, it is very important approaches in biology to understand the history and to explain the reasons of the incongruence in the tree of life. Several reports claimed new algorithms to solve the previous mentioned problem. One of the commonly used methods in these issues is tree comparison. The other used method is distance matrix phylogenetic profiling method.

The tree comparison method is proposed to be technique to detect and quantify the HGT in the species based on the anomalies in phylogeny and incongruence of trees in prokaryotes (Brown, 2003; Creevey, et al., 2004; Eisen, 2000; Garcia-Vallve, et al., 2003; Koonin, et al., 2001; Philippe and Douady, 2003; Puigbo, et al., 2007; Ragan, 2001).

Several methods have been claimed to be used as techniques in the tree comparison, like the method to find an agreement tree (Kubicka, et al., 1995; Puigbo, et al., 2007), another method compare the distances based on the number of nodes "Nodal method" (Steel and Penny, 1993), the third method is "Quartets and Triplets" (Estabrook, et al., 1985; Puigbo, et al., 2007), and also the "Split" method (Robinson and Foulds, 1981).

TOPD/FMTS is a software used to compare the phylogenetic tree (Puigbo, et al., 2007). The application consists of two programs; the first is the TOPD (TOPological Distance) program, and the second is the FMTS (From Multiple To Single) program. The TOPD program was the main concern of this study. The previous mentioned comparison algorithms are already integrated with the program. The tree comparison methods that mainly used in this study are the Disagreement, the Nodal, and the split methods due to simplicity and time used to run.

5.1. DISAGREEMENT

At first, Penny and colleague have studied the similarities and abnormalities using tree comparison (Penny and Hendy, 1985), followed by other reports that presented the "Disagreement" method (Kubicka, et al., 1995; Puigbo, et al., 2007).

The method is depending on removing one taxon from each tree and then it compares the two trees together and measure the split distance. The method is repeated until the split distance becomes zero. The method here returns the taxa which if removed the two trees will be identical (Puigbo, et al., 2007) .

5.2. NODAL DISTANCE

The concept of this method is to calculate distances of the taxa to each other by number of nodes in between. On the other word, it counts the number of nodes that separate each taxon to the others.

Then, construct a matrix of each tree. The matrix is a distance (number of nodes) between each taxon and other partners on the tree. The difference between the two matrices calculated, and hence, the Root Mean Squared Distance (RMSD) between the two matrices is defined. The RMSD is the nodal distance between the two trees (Puigbo, et al., 2007).

5.3. SPLIT

The method is dividing the tree topology into parts and then compares the two trees against each other. The final result is the calculation of the (different/possible) ratio.

For example, In case of the algorithm divide the tree into two parts. The possible here is two. The two trees are identical means that the different is 0 and possible is two, and the final result is two. The two trees will be totally different if the distance equal to one (Puigbo, et al., 2007).

CHAPTER 2

CONSERVATIVE MARKERS DETECTING TOOL

1. ABSTRACT

Finding a phylogenetic marker in species plays the important role to uncover the evolutionary relationship of the species. There are numbers of methods for whole-genome comparisons require with many limitations; moreover, bacteriophage genomes cannot be easily analyzed by these methods. This work suggests better approach and aims to provide an automated prediction system of phylogenetic marker.

A new application called "CoMark" has been developed to theoretically predict the conservative proteins over the completed genomes. These predicted proteins are applicant to be phylogenetic marker for the strains that used by the program.

The prediction system was initialized by searching a large collection of whole protein sequences into groups based on their protein sequence similarity. From each individual group, each data set was then aligned and used to determine an evolutionary distance between all pairs of closely phage proteins that calculated the average similarity between genomes. Then the generated phage proteome tree was compared with all tree s which are constructed using each data set. Experiments were conducted with 5797 (5821 including query) protein sequences over 96 (97 including query) bacteriophage genomes.

The proposed system employed proteomic context to predict a genetic marker and could detect marker correspondence in whole-genome comparisons. Although my experimental focus was on bacteriophages, the method might be extended to other microbial genomes as they shared a number of similar characteristics with phage genomes such as functional protein conservation.

2. INTRODUCTION

In the age of the molecular evolution and due to the fast increasing molecular sequence data, there is an increasing concern to construct the evolutionary tree of different organisms. However, the phylogenetic analysis using sequence data is inadequate in case of long branched trees. Furthermore, the branching patterns of any species while constructing the phylogenetic trees is influenced by number of factors and hence affect the results of the phylogenetic analysis; multiple changes at a given site, the long-branch attraction effect and differences in evolutionary rates. These factor may be a reason of the contradictory results or the tree cannot preserve certain evolutionary relationships of the species included in the tree (Bacteriodetes and Spirochaetes; Gupta, 2005).

Here, for more accuracy of the resultant data the one must find not only novel but also more reliable types of motifs. Each one of these motifs is act as a marker that preserves enough information about the evolutionary relationships of these species. This novel marker is then named as a phylogenetic marker (Krauss, et al., 2008). The Phylogenetic marker should have no or small predictable variation within a given species and their sequences are available for most or all species of a genus. These mean that the phylogenetic marker is conservative fragment.

Recently, numbers of methods and algorithms based on sequencing techniques have been developed to study the phylogenetic relationships. One of these techniques is whole-genome comparisons are appeared (Mirkin, et al., 2003; Wolf, et al., 2001); the other method used the gene contents to determine the phylogeny (Snel, et al., 1999), moreover, Genome Blast Distance Phylogeny [GBDP] method (Henz, et al., 2005).

Besides, a new method to construct the phylogenetic tree of the phages called whole-proteome tree have been developed (Rohwer and Edwards, 2002).

On the other hand, one of the most extraordinary phenomena in prokaryotes is the Horizontal gene transfer (HGT) mechanism (Brown, 2003; Doolittle, 1999; Kurland, et al., 2003). The process is enriched in Archaea and bacteria due to the high rate of the Transformation, Transduction, or Bacterial conjugation between bacteria and bacteria or other organisms (Doolittle, 1999; Jiang and Paul, 1998; Kidambi, et al., 1994; Lorenz and Wackernagel, 1994). Both of these species may be ortholog or homolog, this means that gene transfers even to the distant genomes. Here, the genomes of species which composed of genes from multiple sources fail to construct a neither universal nor unique phylogenetic tree. In the other word, the HGT mechanism is claimed by some reports to be the answer the anomalies in phylogeny and incongruence of trees in prokaryotes (Brown, 2003; Creevey, et al., 2004; Eisen, 2000; Garcia-Vallve, et al., 2003; Koonin, et al., 2001; Philippe and Douady, 2003; Puigbo, et al., 2007; Ragan, 2001).

As the importance of HGT, numbers of methods have been developed to detect it. The incongruence of tree is one of the important methods of detection (Brown, 2003; Philippe and Douady, 2003), or some other new procedures try to quantify the HGT (Mirkin, et al., 2003; Philippe and Douady, 2003).

Bacteriophages, viruses that infect bacteria, were first described in the early 1900s (Twort, 1915). It is hardly to classify the phages on basis of genetic markers as they are highly diversified and does not contain conserved motifs like rRNAs; however, structural proteins like capsids could hypothetically serve as a basis for phage taxonomy (Fuhrman, 1999;

Hambly, et al., 2001; Le Marrec, et al., 1997; SUSSKIND and BOTSTEIN, 1978).

In this study, the whole-proteome tree of the bacteriophage strains (Rohwer and Edwards, 2002) and then compared this whole-proteome tree with other dataset trees using TOPD (TOPological Distance) tree comparison program (Puigbo, et al., 2007). By using methods such as 'Nodal', 'Split' and 'Disagreement' methods built in the program and based on the incongruence of the trees compared, one can identify the horizontally transferred proteins from the other proteins. Then, one can exclude the most conservative proteins among the proteins of these used species.

The motivation of this study is to present a precise and user friendly program to theoretically find the most common and conservative protein for species with complex proteome and high rate of gene transfer. In this study, the bacteriophage strains have been used as an example strains due to their complicated genome structure and due to their importance in the atmosphere.

3. METHODOLOGY AND ALGORITHM

DATASET

At the beginning, 96 phage species including 6 strains has been chosen; these species are belonging to Lactobacillus phage, Listeria phage, Streptococcus phage, Bacillus phage, Lactococcus phage, and Staphylococcus phage strains; the phage names are according to the host. The sequences have been downloaded from the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=6&name=Phages).

The total number of protein sequences is 5797 sequence come from 96 phage strain. Then, the query species EFAP-1 also contains 24 protein sequence has been added to the database. The final database used is consisting of 5821 protein sequence.

Table 2-1 and Appendix-1 contain the full list of all strains that used in the study and the number of the protein sequences belongs to each strain.

Table 2-1. The table shows phage species with number of strains (Strains #) used and number of protein sequences (Proteins#) belongs to each strain. The total number of strains is 97, while the total number of proteins is 5797.

Phage Species	Strains #	Proteins #	Strains used
EFAP-1 phage	1	24	EFAP-1
Listeria phage	7	457	2389, A006, A118, A500, B025, B054, and P35
Bacillus phage	9	688	BCJA1c, Cherry, Fah, Gamma, IEBH, phi105, SPBc2, SPP1, and WBeta
Lactobacillus phage	10	542	A2, KC5a, Lc-Nu, LL-H, Lrm1, phiadh, phiAT3, phig1e, Lactobacillus prophage Lj928, and Lactobacillus prophage Lj965
Streptococcus phage	12	559	2972, 7201, 858, DT1, MM1, O1205, phi3396, Sfi11, Sfi19, Sfi21, SM1, and Streptococcus thermophilus bacteriophage Sfi11
Lactococcus phage	21	1086	1706, bIBB29, bIL170, bIL285, bIL286, bIL309, bIL310, bIL311, bIL312, BK5-T, c2, jj50, P008, P335 sensu lato, phiLC3, Q54, r1t, sk1, TP901-1, Tuc2009, and ul36
Staphylococcus phage	37	2465	11, 187, 2638A, 29, 37, 3A, 42E, 47, 52A, 53, 55, 69, 71, 77,

			80alpha, 85, 88, 92, 96, CNPH82, EW, PH15, phi 12, phi13, phiETA, phiETA2, phiETA3, phiMR11, phiMR25, phiN315, phiNM, phiPVL108, phiSLT, PVL, X2, Staphylococcus prophage phiPV83, and Staphylococcus aureus phage phiNM3
TOTAL	**97**	**5821**	

PROCEDURE

PRIMARY STEP

The database file is constructed of 5797 protein sequences which come from 96 bacteriophage genomes as a primary database. Then, the query species contains 24 protein sequences more has been added to this database. The whole database used is 97 strains and 5821 protein sequences.

The procedures of constructing the whole proteome tree are described by former research done by F. Rohwer (Rohwer and Edwards, 2002). Figure 2-1 is schematic diagram to illustrate the general steps of the procedure. Figure 2-2 is a work flow chart of the steps of the procedure used. Figure 2-3 is a pseudo-code illustrates the algorithm used in the current study.

The database file (contains 5821 protein sequences) were compared against the protein sequences of the other viruses using NCBI’s Basic Local Alignment Search Tool (BLASTP) program (Altschul, et al., 1990; Altschul, et al., 1997) to calculate the BLASTP distances and BLOSUM62 matrix, with cutoff E- value used is <0.1. The BLASTP output file resulted from this step is a result of comparison between all sequences against each other. The file is a group of multiple protein sequences aligned against each other. These sequences are belonging to different species in addition to the protein sequences of the query species or species under test. In many cases, more than one protein sequences from only one species aligned together with different E-values for each.

However, the initial database input file employed in this method is constructed of proteome sequences from bacteriophages strains; one can use any other proteome sequences from any other organism.

Here, the study of bacteriophages and EFAP-1 phage used as a query species (test species) was the main focus of the current study.

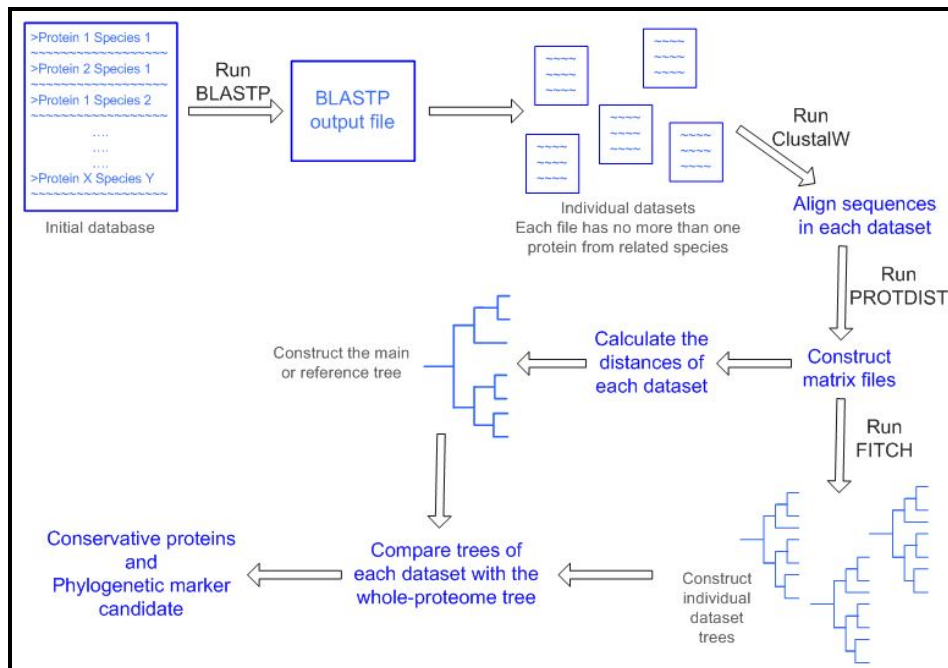


Figure 2-1. Schematic diagram shows the procedures and main steps of the algorithm.

TREE CONSTRUCTING STEP

The next step is to eliminate the duplicated proteins come from the same species. Only the protein with highest e-value from each species is being excluded. The protein sequences obtained together and written to new FASTA file format. Each new file is considered as a dataset file for the next step.

Each dataset aligned by CLUSTALW program (Thompson and Higgins, 1994) to get Phylip format files; then the Phylip formatted files passed to PROTDIST program (Felsenstein, 1990) to calculate the protein distance scores of each files. The default setting of the program has been used, in addition to Dayhoff-PAM matrix option. These proteomic distances datasets are used to construct the trees of each datasets using FITCH program (Felsenstein, 1990); here, the default setting of the program has been used as well. The final output of this step is number of trees each one belongs to one of the datasets. These trees can be considered as the “individual trees” of the datasets.

MATRIX CONSTRUCTING STEP

On the other hand, the datasets of protein distance scores created by PROTDIST program are employed to new matrix of whole-proteome tree of strains used. The calculation of average distances and penalties for absent genome are fully described in Rohwer paper (Rohwer and Edwards, 2002). This is the preliminary step to construct the whole-proteome tree

for phage. In this study, this tree is considered as a "whole tree", "reference tree" or "main tree" that consists of 97 phage strains.

TREE COMPARISON STEP

In this step, and using the tree comparison software TOPD (TOPological Distance)/FMTS (From Multiple To Single) [TOPD/FMTS] program (Puigbo, et al., 2007), the individual trees are ready to be compared against the main or reference tree. The Nodal, Split, and Disagree algorithms are built in methods in TOPD/FMTS program; in this paper, the three methods were the major tree comparing methods. The program parameters have been adjusted to be non-random comparison and multiple reference tree option.

CONSERVATIVE PROTEIN FINDING STEP

The main goal of this step is to determine whether the protein is horizontally transferred or vertically transferred protein.

Here, by comparing between the whole-proteome tree (reference tree) and the individual trees using Disagreement, Nodal, and Split methods built in TOPD/FTMS. Among the results obtained, the smallest and closest 1% scores resulted from these three comparisons have been extracted.

The dataset which found in the results of these three comparison methods are known here as a "conservative protein".

Other than the triple existence of datasets, the datasets scored at least with two methods (double existence) as well as the datasets scored with only one method out of the three methods (single existence) have been determined.

Figure 2-2 is flow chart of the algorithm used by the application. Figure 2-3 is a pseudo-code to illustrate all the steps used in this procedure.

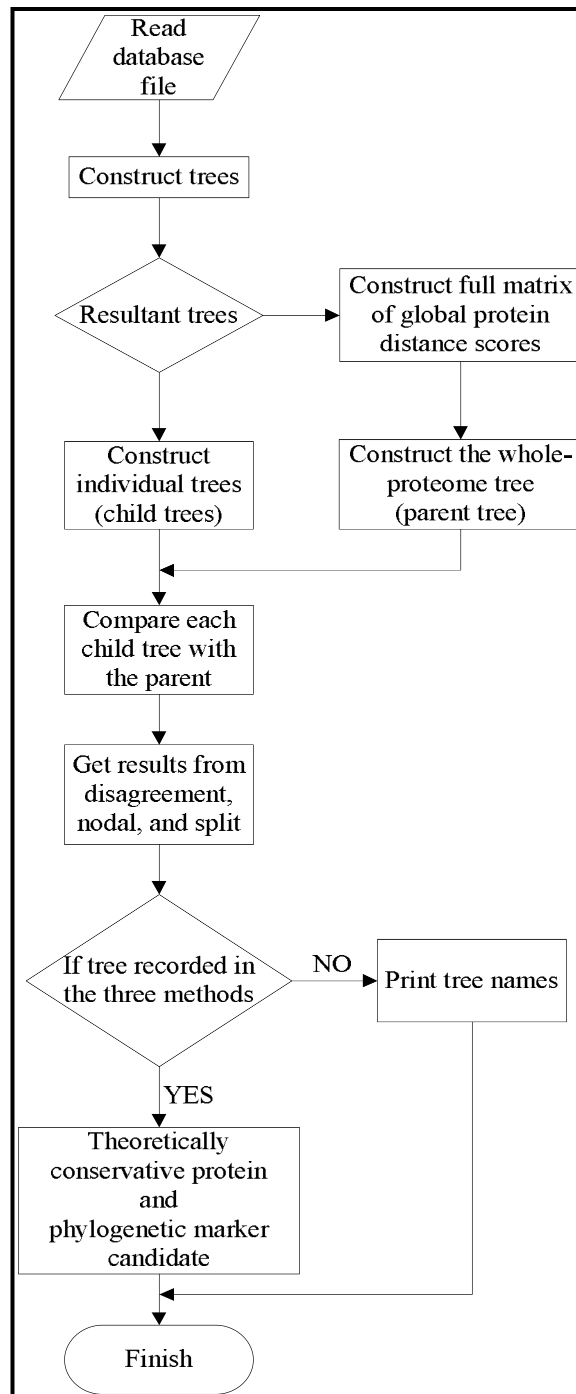


Figure 2-2. The flow chart represents of the algorithm used by the program.

```
1.  START
2.  Read database file
3.  Select the desired strains
4.  Run BlastP program
        a.  BlastP output is query protein and its related proteins
5.  Each query protein and its related proteins collected together to make dataset
6.  FOR each dataset
        a.  Run ClustalW program, to align sequences
        b.  Run Protdist program, to make distance matrix
        c.  Construct new dataset matrix contains all species
        d.  If distance between two species already exist
                i.  Print the distance
        e.  Else
                i.  Print distance penalty score = 100
        f.  ENDIF
        g.  Draw tree (individual tree)
7.  ENDFOR
8.  FOR each dataset
        a.  Construct matrix of the average of distances of all datasets
        b.  Draw the whole genome tree (reference tree)
9.  ENDFOR
10. FOR each tree of dataset
        a.  Compare the individual trees with the reference tree
        b.  Calculate tree comparison distance
        c.  Compare comparisons results against each other
11. ENDFOR
12. Get the lowest 1% of the result
13. Compare the result of each method with other two methods
14. IF the tree recorded in three methods (triple occurrence)
        a.  Print conservative protein in the output file
15. ELSE IF tree recorded in two methods only (double occurrence)
        a.  Print tree name
16. IF tree recorded in only one methods (single occurrence)
        a.  Print tree name
17. END IF
18. END
```

Figure 2-3. Pseudo-code to illustrate all the steps used in this procedure.

4. RESULTS

After ProtDist program ran 4799 matrices have been resulted. Some matrices contains (-1.0) distance scores. This indicates that the finite distance, and both proteins sequences are too different. The final number of the trees obtained with correct distances was 4949 trees.

These trees compared against the main whole tree using the tree methods. Then the smallest distances scores resulted from these three methods are compared to check the duplicate and triplet occurrence of these trees distances scores. The triplet occurrence indicate that this tree or this dataset is may contain a certain set of proteins; one of these proteins is a candidate to be a conservative protein. Here, this conservative protein, which in turn, can be considered as a candidate to be phylogenetic marker. See table 2-2 for the resulted distance relevant to each method used in the study.

Table 2-2. Distances resulted after the trees comparison using the three methods in TOPD program. Each column contains the Tree IDs that belongs to the lowest distance scores corresponding to each method have been used. The total number of trees is 4749. The lowest are 48 trees. In case of, there is one more score close to the last score; the closest scores to this last one have been chosen. The Tree ID starts by "T" then followed by the ID number.

	Disagree		**Nodal**		**Split**	
ID	**Tree ID**	**Distances**	**Tree ID**	**Distances**	**Tree ID**	**Distances**
1	T1049	0.690722	T1391	10.68057	T1049	0.765957
2	T1149	0.773196	T1614	10.30439	T1184	0.851064
3	T1275	0.649485	T1777	10.13277	T1275	0.787234
4	T1391	0.670103	T1909	10.13277	T1391	0.776596
5	T1420	0.71134	T2291	10.13277	T1420	0.840426
6	T1558	0.670103	T2303	8.072986	T1558	0.755319
7	T1685	0.690722	T2330	8.090035	T1685	0.765957
8	T176	0.680412	T236	8.097704	T176	0.787234
9	T18	0.659794	T2402	10.14509	T18	0.765957
10	T1850	0.721649	T2526	8.004348	T1850	0.808511
11	T1900	0.701031	T282	10.13277	T1900	0.797872

12	T2303	0.690722	T293	7.91118	T2303	0.765957
13	T2330	0.701031	T3109	9.255478	T2330	0.744681
14	T236	0.628866	T3399	10.6389	T236	0.712766
15	T2526	0.701031	T3533	10.38225	T2526	0.744681
16	T2761	0.690722	T3599	10.68057	T2761	0.755319
17	T2833	0.680412	T3670	9.747367	T2833	0.787234
18	T2902	0.690722	T379	10.43959	T2902	0.755319
19	T293	0.670103	T3792	10.38225	T293	0.755319
20	T2971	0.721649	T3857	10.38225	T2971	0.776596
21	T3039	0.690722	T3926	9.447961	T3039	0.765957
22	T3109	0.752577	T3929	10.14509	T3109	0.797872
23	T317	0.670103	T3938	10.75745	T3150	0.851064
24	T3343	0.639175	T3979	9.447961	T317	0.765957
25	T3399	0.71134	T3990	10.75745	T3225	0.851064
26	T3465	0.680412	T4005	10.14509	T3287	0.851064
27	T3533	0.639175	T4054	8.503474	T3343	0.776596
28	T3599	0.670103	T4055	10.54828	T3399	0.776596
29	T3670	0.649485	T4069	10.75745	T3465	0.734043

30	T3735	0.690722	T4080	9.447961	T3533	0.734043
31	T379	0.690722	T4118	10.14509	T3599	0.776596
32	T3792	0.639175	T4120	10.75745	T3670	0.744681
33	T3857	0.639175	T4130	9.447961	T3735	0.744681
34	T3926	0.762887	T4168	10.14509	T379	0.734043
35	T3979	0.762887	T4178	10.44101	T3792	0.734043
36	T4054	0.690722	T4215	10.42586	T3857	0.734043
37	T4080	0.762887	T4402	10.31955	T3926	0.787234
38	T4130	0.762887	T4413	9.744745	T3979	0.787234
39	T4230	0.731959	T4501	10.50487	T4054	0.765957
40	T4413	0.670103	T4503	10.67729	T4080	0.787234
41	T4628	0.752577	T4541	10.13277	T4130	0.787234
42	T4629	0.649485	T4616	10.13277	T4230	0.787234
43	T4722	0.639175	T4710	10.13277	T4413	0.776596
44	T4749	0.659794	T4722	8.239137	T4628	0.840426
45	T478	0.680412	T4743	10.65975	T4629	0.829787
46	T545	0.649485	T4749	8.86935	T4722	0.797872
47	T89	0.71134	T545	10.11032	T4749	0.765957

48	T939	0.71134	T579	10.66319	T478	0.787234
49	X	X	T751	10.14509	T545	0.744681
50	X	X	T939	7.397162	T637	0.851064
51	X	X	T988	10.14509	T89	0.765957
52	X	X	X	X	T939	0.755319

24 dataset trees found shared in the three methods (triplet occurrence) with low distance scores compared to the other datasets. 23 dataset trees found with low scores with split and disagreement methods but not found with nodal method. No datasets are shared neither between nodal and split nor between nodal and disagreement methods. 27 dataset trees have low nodal distance only, 5 dataset trees have low split distance, while 1 dataset tree has low disagreement distance score. The total number of trees that are either duplicated or found only one time is 56 trees. See the table 2-3 for logical relation of the tree relevant to each method.

Table 2-3. Logical relations of distances resulted from three methods after tree comparison. The triple occurrence of the dataset tree "(split AND nodal AND disagree)"; it means that the dataset distance score of nodal was among the 1% of the lowest scores, and lowest 1% with split and among the lowest 1% scores with disagreement method; 24 trees have been resulted. The "(disagree AND nodal) NOT split" expression means that the dataset has double occurrence with disagree and nodal but not among the lowest 1% when the split method is applied. The same logic applies for "(disagree AND split) NOT nodal" and "(nodal AND split) NOT disagree". Then, the number of datasets scored only single occurrence with any of these methods. The "split OR nodal OR disagree" term refers to the number of trees with other than the triplet occurrence; 56 trees have been resulted.

Logical relation	Number of trees
split ∩ nodal ∩ disagree	24
(disagree ∩ nodal) - split	0
(disagree ∩ split) - nodal	23
(nodal ∩ split) - disagree	0
disagree only	1
nodal only	27
split only	5
split ∪ nodal ∪ disagree	56

The method succeeds to find 24 triple existed datasets, among these datasets; it is assumed that one can find out phylogenetic marker candidate. The table 2-3 has full explanation of this datasets with number of trees and name of dataset tree. The figure 2-4 is a Venn diagram to demonstrate the number of tree belongs to each method and number of trees interrelated between methods together.

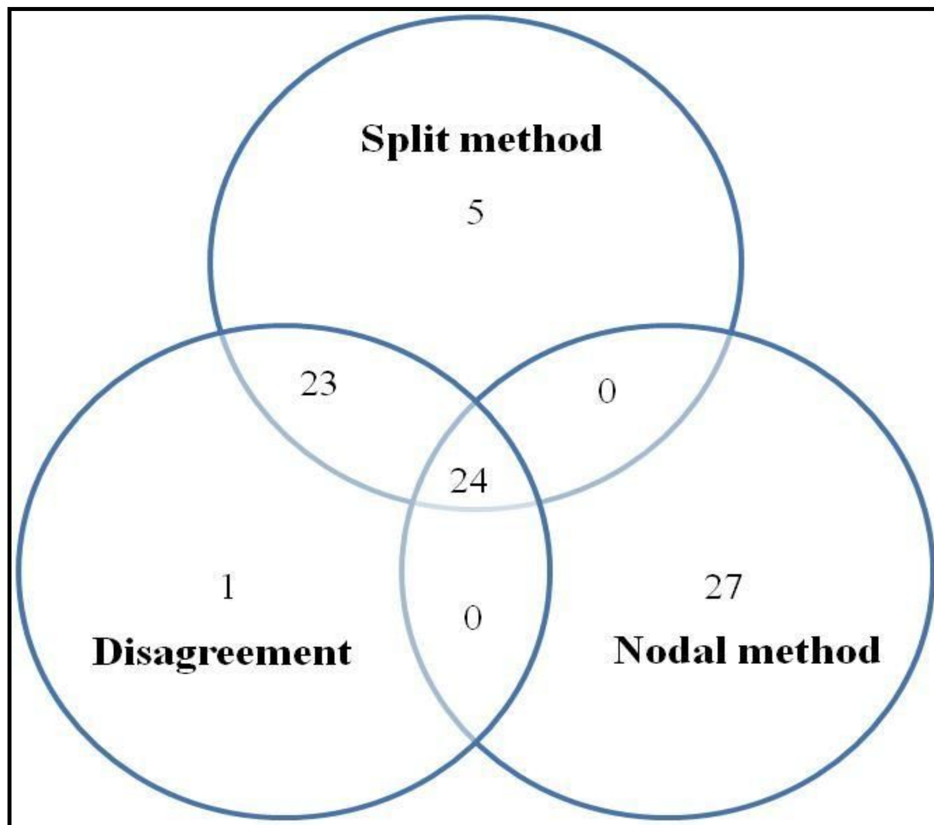


Figure 2-4. Venn diagram represents the resultant distances of highly similar trees using 'Nodal,' 'Split,' and 'Disagreement' methods. The integers represent the number of trees owing to each method. The integers represent the number of trees owing to each method. The intersected areas are the number of datasets scores of more than one method.

First candidate marker dataset included tape measure protein (TMP) and this protein was used to phylogenetic analysis in previous research (Liao, et al., 2008; Pedersen, et al., 2000). It suggested that they have a common ancestor or the same transition in the process of evolution according to the phylogenetic tree based on amino acid sequences of TMPs in phage species which including these proteins. Therefore this protein can be used to determine phylogenetic relationship as well in phage species that used in this study.

Appendix-2 contains the list of conservative proteins (Phylogenetic marker candidate) that belongs to three dataset trees. Each one of these trees appeared to have very low distance scores by using various methods of tree comparison at the same time. Each tree has the list of proteins, these protein sequences are resulted from the first step by using BLASTP and they are related to each other.

Each resultant dataset of protein contains set of proteins which are related to that main protein. The table shows three different datasets which have low scores with the three methods (triple occurrence). These trees contain the conservative and most common proteins among the strains used in the study. In this table, the proteins belonging to that three dataset are almost similar. This leads us to say that these proteins are most common in existence among the strains, less affected by horizontal gene transfer and then these proteins have crucial importance for these strains.

Here, due to the similarity between different datasets, the method purposed in this report is very strong way to find the most common proteins and vital proteins among the strains. These proteins may have a role to be a marker or basis of phylogenetic tree construction.

The other important result is the unknown proteins from the unclassified species are found related in sequences to the other known proteins. This shows that these proteins are structurally or functionally related proteins.

The proteome distance matrix calculated from the summation of average of individual matrices as in Rohwer report (Rohwer and Edwards, 2002). Then, the whole proteome tree constructed with the query species is shown in the Figure 2-5. The tree is display using the MEGA 4.0 program (Tamura, et al., 2007).

As it shown the tree is divided into 7 main clusters, each one corresponds to the strain of phage. The query species (EFAP-1 phage) is displayed in the tree in red box. EFAP-1 is located in Bacillus group and close to the Lactobacillus cluster (Figure 2-5).

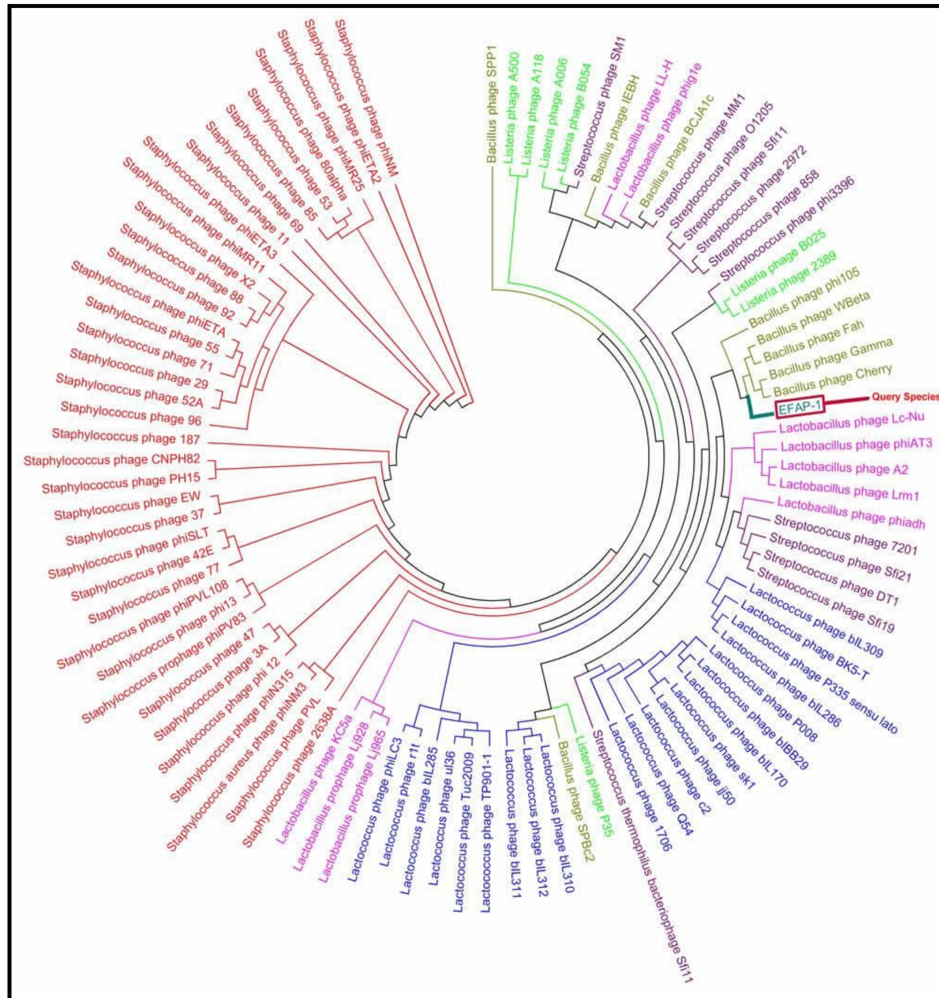


Figure 2-5. The whole-proteome tree of phage strains that used in the method. The red box shows the query species under test “EFAP-1 phage” with relationship with other species in tree. The phage proteomic tree was constructed using a Newick file and displayed using the MEGA 4.0 program. The length of the tree is represented by ‘10’ (proteomic distance score). The proteomic distance score was calculated from the sums of the length-corrected protein distance scores and the penalties and then divided by the total average lengths and the number of missing proteins. Briefly, units used in the phylogenic tree represent values of average amino acid change.

To validate the results, trees based on single files have been constructed and compared the resultant trees agianst the whole-proteome trees, Figure 2-6 and 2-7. EFAP-1 in the whole tree is placed in one cluster with Bacillus group and close to the Lactobacillus cluster. Figure 2-6 is a tree of putative tape measure protein [*Lactobacillus phage Lc-Nu*] and its related proteins; EFAP-1 phage is placed with Lactobacillus phage. In Figure 2-7 is the tree of phage tail tape measure protein [*Staphylococcus aureus phage phiNM3*]; EFAP-1 placed in one group with Listeria phage. The values in the figures represent the robustness of the phylogenetic analysis, which determined by bootstrap analysis (Felsenstein, 1985) with 1,000 replications.

The results obtained from this step are showing that it is hard to find one and clear common phylogenetic tree based on one protein and based on small fragment of the sequence for the phage species that are being used in this study.

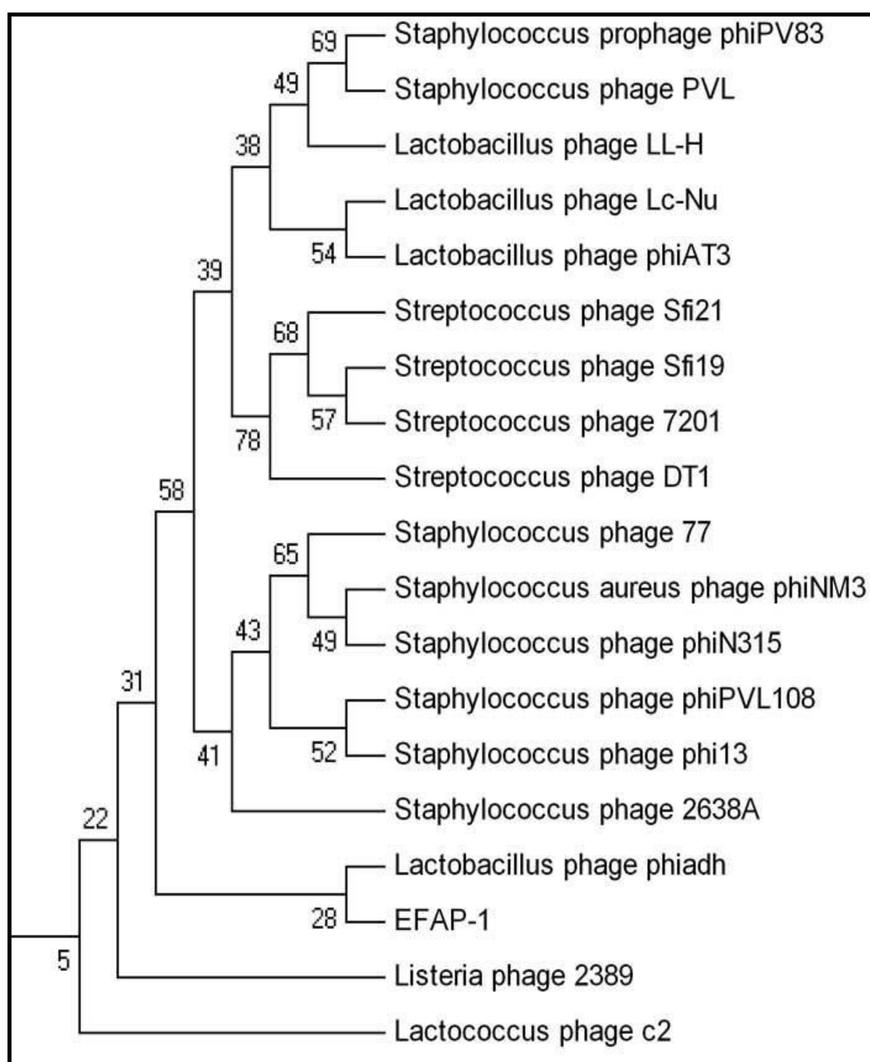


Figure 2-6. Part of the tree constructed of single dataset based on protein putative tape measure. The values represent the robustness of the phylogenetic analysis determined by bootstrap analysis.

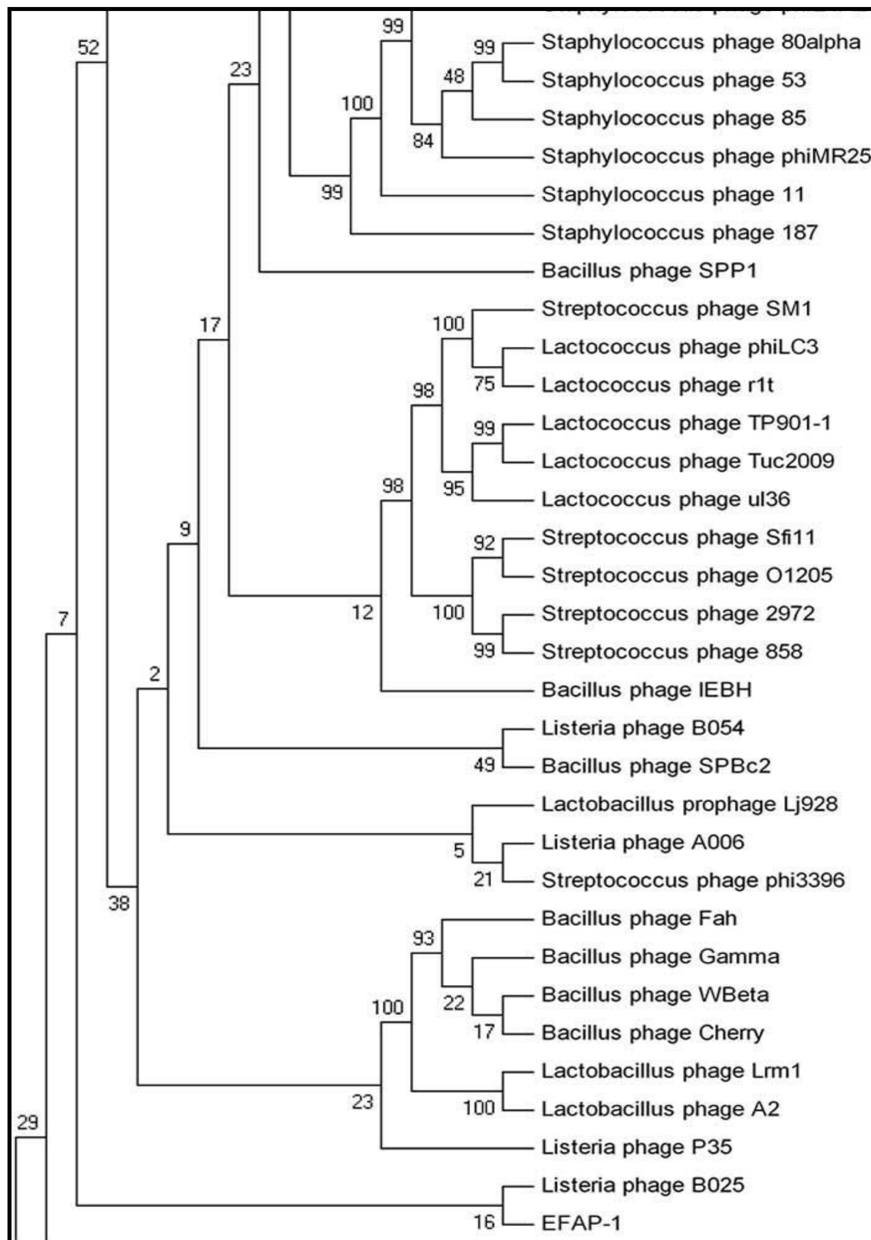


Figure 2-7. Part of the tree constructed of single dataset based on phage tail tape measure protein. The values represent the robustness of the phylogenetic analysis determined by bootstrap analysis.

5. DISCUSSION

From previous report (Puigbo, et al., 2007), it claims that the incongruence in the topology from one tree to another is may be a consequences of Horizontal gene transfer. The nodal distance score is calculated using the root-mean-squared distance (RMSD); if the two trees of the dataset have a RMSD equal 0, this indicates that these are identical or typical tree. The dataset with has Nodal scores is indicating to vertical gene transfer while as RMSD increases this explained as horizontal one. Here, the lowest scores resulted from nodal, split and disagree have been compared against each others to ensure the resulting protein from this step is conservative protein.

It is been reported that absence of single gene which is common between all the phage species and hence used as a basis of classification system (Rohwer and Edwards, 2002). This is the reason that ICTV taxonomy system is based on structural proteins like capsids as a basis of classifications. In the study of Rohwer and his colleague, they proposed the method which is compatible with the ICTV systems and to overcome the problems of the conventional methods of the taxonomy.

From the above it was very hard to find neither conservative proteins nor phylogenetic markers among the phage species based on single or small dataset of proteins.

The results obtained here is compatible with the results of other reports (Rohwer and Edwards, 2002) and (Puigbo, et al., 2007). Emerge between both methods will aid to construct highly precisely tree of phages and to find theoretical vertically gene transferred protein (conservative).

The method is depending on a large dataset of proteins, and each strain is expressed by multiple proteins with different length. In the other hand, since the results are covering long range of strains proteome, the results here are more realistic than using the small databases of proteins.

The method succeeds to construct the phylogenetic tree, show the most common proteins and less affect with horizontal gene transfer proteins. Moreover, the method is placed the un-annotated proteins together with the one of a known annotation.

5.1. THE ALGORITHM APPLICATIONS

The HGT process has lot applications in nature or even in the industry. The quantification, and detection of the HGT will help the further research improvements.

There is number of the important applications of HGT in biology can be summarized as the following:

1. The HGT may play a role in the genomic alteration, changes in the evolutionary phylogenetic relationships, and the tree incongruence.

2. The changes in coding or non-coding regions yield unexpected proteomic products. Therefore, the protein-protein interactions and network may also be affected.

3. The HGT has a direct or indirect relationship with the pathogenicity of the microbes. This could be positively by acquiring a desired genes, or negatively when a virulent genes being transferred to another species.

4. One of the most important implications in the biomedical field is the transfer of the antimicrobial (antibiotic) resistance genes between species.

5. One of the most important implications of HGT is the industrial approaches, like biotechnological and diary industries that depend on the bacteria or protein products derived from bacteria.

6. The plant breeding, genetically modified plants (GMP) and biotechnological crops are one of applications of the artificial HGT.

7. The production of the transgenic animals is also produced by introducing one or more DNA into embryonic stem cells.

8. The discovery of a new generation of vaccines and drugs is one of the most useful impacts of HGT process.

The application is studying the phylogeny and proteome of species with high rate of mutations. This detection of the HGT in these species was the core element of this study. The algorithm will be beneficial in number of cases, and with more improvement will catch the attention of other research in other disciplines. Here the algorithm may be useful in some areas like:

1. The main aim of this study is the phylogenetic and evolutionary study and to track the history of a certain proteins.

2. This approach will help to identify a phylogenetic marker for some unclassified species and the species whose genome is a complicated one.

3. The algorithm succeeds to identify a common protein by using a simple database of proteome of certain species.

4. The study of phylogenetic markers, common, and conservative proteins is very crucial to the field of biology. These approaches will solve many problems in the metagenomic era and the tree of life.

5. The method used the whole-proteome dataset of a punch of species, which an important research pipeline.

6. An important trend in the phylogenetic studies is to track the history of the genes. Using the tree comparison method, the study is able to track the most common protein and it is identified as a most conservative protein.

7. One of the most applications of the study is to refine the genes and eliminate the horizontally transferred genes and present the most conservative proteins.

8. The future improvements, like integrating additional functionalities, tree constructing techniques, detect/quantify HGT and/or detecting the gene order, are interesting research pipeline in the field of phylogenetic.

9. The study with some improvements will help the biotechnological researchers.

10. The software is user-friendly, and freeware with full online documentations.

5.2. FUTURE DIRECTIONS AND LIMITATIONS

1. The program is desktop application for the windows systems only. The future approach is to launch a web-based tool to do the same functionalities.

2. The conservative proteins are not necessary a good phylogenetic marker.

3. The whole-proteome tree step needs further improvement. The whole-proteome tree may not yield a correct phylogenetic tree for the species.

4. The FITCH program has been employed to construct the whole proteome tree and other trees. FITCH program is based on distance algorithm. The maximum likelihood or Bayesian algorithms may yield some more and accurate results. The new prospective is to add some more program other than FITCH or PROTDIST programs. This will allow the users to select and switch between several techniques simply within a user-friendly interface.

5. Beside the whole-proteome tree comparison, the single dataset comparisons analysis may be an interesting research line to discover the proteins history.

6. More research efforts should be directed to validate the robustness of the usage of the whole-proteome dataset.

7. In this study best-blast method has been used to eliminate paralogs and detect orthologs. In some cases, this method may not be very robust technique. The proposed improvement will concentrate on including methods such as reciprocal-best-blast (RBH), Markov clustering

(orthoMCL), and/or Inparanoid techniques to give best and more reliable results.

8. The further improvement will include changing the blast threshold.

9. The random test of tree comparison will lead to more accurate results. Some other consideration during the tree comparison steps should be put into the account, since the Nodal distance method are not normalized.

10. The study of the gene order for those genes selected as phylogenetic markers is an interesting and future research line.

6. CONCLUSION

Other programs have been done in field of Bioinformatics to understand the genome structure or to align and construct the phylogenetic tree. Some other software is used to compare the phylogenetic trees. In this program, both methods are combining together in one user friendly program that allows users to use it easily.

The program might help the user to compare the proteomic trees and search for phylogenetic marker inside this tree, as well as the program will place any unknown protein sequences with the known one. The last application is useful in the phylogenetic studies.

In the future, these studies may improve current methods of detecting and quantifying HGT as well as help in developing useful computational algorithms for constructing phylogenetic relationships among species with complicated genome structures.

CHAPTER 3

APPLICATION FEATURES

AND USER MANUAL

1. INTRODUCTION

By entering the age of metagenomics, it is necessary to understand the evolutionary relationships between different organisms. In some cases, it is very hard to construct these evolutionary phylogenetic relationships for some organisms. Here, the role of phylogenetic markers [PM] is that it contains the key information of the phylogenetic relationship. PM is a conservative motif which presents in most of the species within the genus and it has no or small predictable variation among them. The PMs have crucial role in the phylogenetic reconstructions.

Bacteriophages for example lack of the clear genome based classifications. This is because the absence of the common gene along over the all phage species. Thus, Bacteriophages have no common or clear phylogenetic marker.

The conservative markers finding tool is an algorithm to find theoretically most closely related amino acids from different homologous species. These related amino acids seem to be conservative. These conservative proteins are good candidate to be to find the Phylogenetic markers.

2. APPLICATION FEATURES

ABOUT THE PROGRAM

The program is desktop application based on DOT NET framework 2.0, and developed using C#.NET 2005 version.

The program is freeware and copy righted to the author. The program can be downloaded from http://snugenome.snu.ac.kr/comark/.

SYSTEM REQUIREMENTS (PREREQUISITES)

- Windows (x86 32-bit or x64-bit based systems).
- Microsoft .NET Framework Version 2.0 Redistributable Package. (http://www.microsoft.com/downloads/details.aspx?FamilyID=0856EACB-4362-4B0D-8EDD-AAB15C5E04F5&displaylang=en)
- Binary distributions of Perl. (http://www.perl.org/), the Perl language is required to run TOPD Perl script.
- Hard disk space (C Partition) more than 1 Gigabyte (1 GB).
- BLASTP, ClustalW, ProtDist and Fitch [the two programs are part of PHYLIP package], TOPD and MEGA 4.0 programs are integrated in this application and installed automatically with the software.

ALGORITHM

The algorithm and main idea are fully described in chapter 2. Figure 2-1 in chapter 2 is schematic diagram to illustrate the general steps of the procedure. However, this chapter contains a description of the customization of the program functionality to fit the algorithm, and description of main elements of the software.

PRIMARY STEP

The input file format should be in FASTA file format which contains only the name of the proteins and then the sequences. The proteome of any species can be loaded, whether bacteriophage or any other viral or bacterial proteome database file. The figure below shows the standard input format.

Figure 3-1 shows the standard format of the database file used as an input in the application. The input file should contain both the database and the query species in the same file.

```
>gi|orf21| [EFAP-1]↓
MGRPRKLINAQVGNLTTEQQEERKKEEEALYNYEKLDFSYYPQGLLQGAFPEWERISHFIGDLPISELDQQTMVRYCNYT↓
YLYSEATERLMEEGEITPDGKKNPRVDIMNSYSKELKSATADLGLTINARLKIVAPAEKTKESNDPLGQLIKLRQQS↓
>gi|orf22| [EFAP-1]↓
MVKDLHTQQTRAFKSQTEADKFYNKKSGYFKDVRTKLGGRNRHYEIIEVV↓
>gi|orf23| [EFAP-1]↓
MPKRRCKVAHCREYVDIPEVYCEEHKGNTQRTYNKQVRHSPDNKKYADFYASTQWRNVRARKLSMNPMCEVCNASIATIV↓
HHRQEVRTTMGWEHRLDIDNLESICQECHNKEEHSASFRHRKG↓
>gi|orf24| [EFAP-1]↓
MTMEGMTDKQLLSHTKLELVSYIRQLEQAMVTNTSDTTEQPTTEPTSKPKVSYDILSTTNNNKDMWGF↓
>gi|56694870|ref|YP_164380.1| hypothetical protein BCBBV1cgp1 [Bacillus phage BCJA1c]↓
MQLGKCEKCGNTGRAKNLTLYKEEMRCPNCLNLHFMFSYWEGPKRMMQVMDFGEGRMRQIAFDEIEEEDERVRKYIDHHW↓
EKINSKPYEYNPK↓
>gi|56694871|ref|YP_164381.1| hypothetical protein BCBBV1cgp2 [Bacillus phage BCJA1c]↓
MREAERVKKYILIIVATALMLVGCNNNSEAKDIAESYMDSVRNGEDFELIITSEYKFIDVFEYDYLRTLDEVRREDSLEF↓
SREVYDLFRELGEYEEHPTYEDHKEAMKVEFHDHEILEDDNDTLILWNRDGWVEDHTLLYDVVVADEEGNKIYKKAEIIV↓
SFLEGEDRQSEFIRSIKLR↓
>gi|56694872|ref|YP_164382.1| integrase [Bacillus phage BCJA1c]↓
MASYQKRGKTWQYTVSRTVDGKSQPIRKGGFKTKKEAQAAAAEVEAELRKGVTPHLKLEPFDEYFESWMKLYKSDVTNNT↓
KERYKNTLETIKNYFGSKPIQQINKRTYQAFLNEYARTHARATTKKLNTHIRACVKDAIDEGVIRVDFTRKASISGDNKA↓
KKDDEKHLHYRDSQKLLQAVYERLDKSLTHYIILLALTSGMRFAEIVGLKREDFDFKQNTISINKTWGYTKKMHEGFGPT↓
KNEQSVRKIKMDVQTMKAFNNLFDMLPENIHQLVFYSPSSKYKVISNGNANKLLKSLLGELKIEPISMHGLRHTHASVLL↓
YKGVSIYYVSERLGHADIETTMNEYAHIVKELRTQDEEKTAAVFKNMVV↓
>gi|56694873|ref|YP_164383.1| hypothetical protein BCBBV1cgp4 [Bacillus phage BCJA1c]↓
MYVKKAVNQLILKYNTTCPFLLAKQLHIEIEFVNLGRKMLGFYTKNRRVPIITINESVDYMQQTFICAHELGHHELHPNI↓
NTPFLTKNTFYSIDKYEIEAHTFAIELLFANKKIITASDLEVYGIPKQIALLRKYG↓
>gi|56694874|ref|YP_164384.1| repressor [Bacillus phage BCJA1c]↓
MSLVKKIKLMCDEKKITVAELERRVGISNGQIRKWDSSTPGVDKLARVADYFNVSTDYLLGRTDKKRYYELTEKDERDIQ↓
KELEAMINGLNSKDGYAAFDGIDLENMDEEDRELLISALENSLRVAKRVAKQKFTPKKYRD↓
>gi|56694875|ref|YP_164385.1| cro [Bacillus phage BCJA1c]↓
MSEELGIKVRSELFKRKMSQRELAKMIGISEAYLSDIINGRKTGAKAQGHIKHIRKVLSI↓
>gi|56694876|ref|YP_164386.1| hypothetical protein BCBBV1cgp8 [Bacillus phage BCJA1c]↓
MTFGEMAGYTKRFHEIMQLDDQQLRSERLGNLMSDMEVRYQIPIFNKADFNQKNPHVFYLYRAVSQARSL↓
>gi|56694877|ref|YP_164387.1| hypothetical protein BCBBV1cgp9 [Bacillus phage BCJA1c]↓
MKISQMIKQLQDIKEEHGDLDLYEYSDIATIIKRENAYLPRVDKIHYKKHADYPSLKDELHNEDLSVSDENIYDIDLSRP↓
IVKGVVI↓
>gi|56694878|ref|YP_164388.1| hypothetical protein BCBBV1cgp10 [Bacillus phage BCJA1c]↓
MHITKMERRRLKMEFQVGDWVRINHYGNEWTTKVKEVRGRTVIKPEYDIHVSAGSWYHISRVRKATKEEIESTFPRRHHM↓
NNKATLILDRTETSVILDALRFARKEYEPGIIDMVSKKVKPGVISLDGGELKTIYETLFLKGYWCSQQNWTKVTREIWDI↓
GNDVLAVRQAFHKQWEKQLLGRR↓
```

Figure 3-1. Screenshot of the database file format, this is the input database file used by the application (standard FASTA format).

After the database file is loaded, and proceed in the program. Then by clicking construct tree button will fire an event to run NCBI-BLAST (Basic Local Alignment Search Tool) program (Altschul, et al., 1990; Altschul, et al., 1997). The program can be downloaded from National Center for Biotechnology Information (NCBI) website. First step is to format the database file using Formatdb program (one of the BLAST program package).

Then the BLASTP (Standard protein-protein BLAST) program is ran automatically to align the related protein sequences together using BLOSUM62 matrix, with cutoff E- value used is <0.1. The BLASTP output file resulted from this step is a result of comparison between all sequences against each other. The file is a group of multiple protein sequences aligned against each other. These sequences are belonging to different species in addition to the protein sequences of the query species or species under test. In many cases, more than one protein sequences from only one species aligned together with different E-values for each. Figure 3-2 represents the output file resulted from BLASTP program.

```
Query= gi|78000025|ref|YP_358771.1| putative tape measure protein
[Lactobacillus phage Lc-Nu]
         (1522 letters)

Database: C:/PTC_Files/My_Query_File.faa
           5821 sequences; 1,175,404 total letters

Searching..................................................done



                                                                 Score    E
Sequences producing significant alignments:                      (bits) Value

gi|78000025|ref|YP_358771.1| putative tape measure protein [Lact...  2924   0.0
gi|48697271|ref|YP_025038.1| putative minor tail protein [Lactob...  1069   0.0
gi|29165636|ref|NP_049403.2| putative tail component protein [St...   458   e-130
gi|9632950|ref|NP_049978.1| putative minor tail protein [Strepto...   452   e-128
gi|9632906|ref|NP_049935.1| minor tail protein [Streptococcus ph...   446   e-126
gi|9634660|ref|NP_038334.1| ORF33 [Streptococcus phage 7201]          429   e-121
gi|13095855|ref|NP_076745.1| tail protein [Lactococcus phage bIL...   283   4e-077
gi|9633052|ref|NP_050160.1| hypothetical protein phiadhp52 [Lact...   259   7e-070
gi|148750854|ref|YP_001285897.1| hypothetical protein LPLLH_ORF3...   215   1e-056
gi|23455811|ref|NP_695158.1| minor capsid protein [Lactobacillus...   204   3e-053
gi|9633051|ref|NP_050159.1| hypothetical protein phiadhp51 [Lact...   196   6e-051
gi|13095856|ref|NP_076746.1| tail protein [Lactococcus phage bIL...   155   1e-038
gi|41179260|ref|NP_958593.1| putative putative minor tail protei...   147   3e-036
gi|orf8| [EFAP-1]                                                     132   2e-031
gi|13095795|ref|NP_076686.1| tail protein [Lactococcus phage bIL...   127   3e-030
gi|157325231|ref|YP_001468653.1| Tmp [Listeria phage B025]            124   4e-029
gi|14251140|ref|NP_116507.1| unknown [Lactococcus phage BK5-T]        122   1e-028
gi|30089904|ref|NP_839934.1| putative tail lysin [Lactococcus ph...   121   2e-028
gi|29028708|ref|NP_803396.1| tail length tape measure protein [S...    99   1e-021
gi|119443702|ref|YP_918940.1| tail length tape measure protein2 ...    97   4e-021
gi|9635181|ref|NP_058455.1| hypothetical protein PVL_16 [Staphyl...    95   2e-020
gi|9635728|ref|NP_061641.1| phi PVL ORF 17 homologue [Staphyloco...    95   3e-020
gi|29028657|ref|NP_803346.1| tail fiber protein [Staphylococcus ...    92   2e-019
gi|66395509|ref|YP_239871.1| ORF001 [Staphylococcus phage 42E]         92   2e-019
```

Figure 3-2. Screenshot of the output file resulted from BLASTP program.

TREE CONSTRUCTING STEP

Here, the program eliminates the duplicate of proteins come from the same species. The protein sequences obtained together and written to new fasta file format. Each new file is considered as a dataset file for the new step.

Each dataset aligned by CLUSTALW program to get phylip format files; then, files passed to PROTDIST program to calculate the protein distance scores of each files. These proteomic distances datasets are used to construct the trees of each datasets using FITCH program.

MATRIX CONSTRUCTING STEP

After the alignment of sequences against each other, the resulted file is ready to purify the data as mentioned above. The new purification data is to remove the multiple existences of proteins which are belonging to the same species. The reason of this step is to avoid data redundancy.

The new dataset files are containing the protein sequences of only one protein from each related species. Each dataset is FASTA format and ready for alignment.

The program automatically will run PROTDIST and calculate the protein distance scores and then run FITCH program to construct trees of each dataset.

It is important to notice that the new matrix of each dataset represent a matrix of all species that found in the initial database file. Here, the species that are originally present in each matrix file have been added as

well as the other species that are not found in this dataset. The new species added with 100 penalty scores. The aim of this step is to that the new tree constructed of each dataset will be composed of the all species. Therefore, it eases the comparison between the trees of each dataset against the whole-proteome tree (the main tree).

TREE COMPARISON STEP

The program uses the tree comparison software TOPD/FMTS, the individual trees are ready to be compared against the main or reference tree. The methods used are Nodal, Split, and Disagree algorithms (built in TOPD/FMTS program); with options of NON-random comparison and multiple reference tree option.

CONSERVATIVE PROTEIN FINDING STEP

The "tree comparison" button will initiate the tree comparison step and construct the matrix. The last option is to view the tree of the whole genome as well.

The step is to determine the protein whether is horizontally transferred or vertically transferred. First, by comparing the whole-proteome tree (reference tree) against individual trees using TOPD/FTMS. Among the results obtained, the smallest 1% scores resulted from these three comparisons have been extracted.

The dataset which found in the results of these three comparison methods are here known as a "conservative protein".

Other than triple existence of the results the double existence and the dataset who has score in only one method of these three methods are determined and printed in the result files.

3. PROGRAM OVERVIEW

UPLOAD NEW DATABASE FILE WINDOW

The window is used to upload new database file to the program memory and start a new query process. The input format is FASTA format and as it shown in the figure 3-1. The window can be reached by clicking “File”, then click on the “New Database” button. Figure 3-3 is a screenshot of the window of the program.

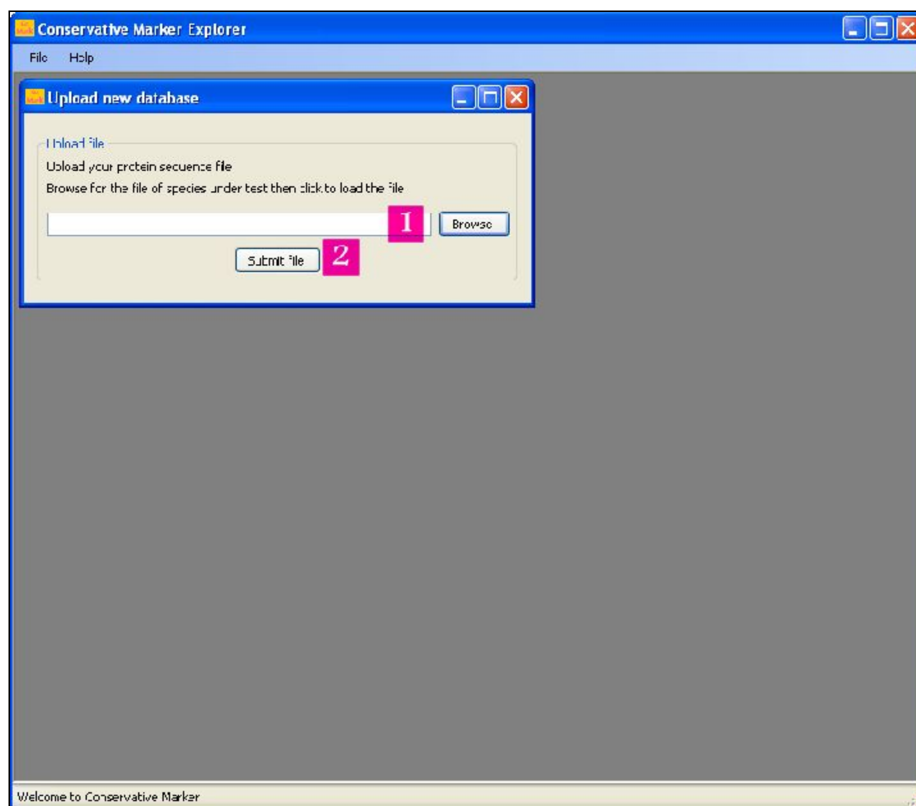


Figure 3-3. The screenshot shows the main interface of the program. The window is to load a new database file into the program memory. The database file then will be parsed and ready for processing. The user can get this window by clicking "**File"** in the toolbar, then, clicking "**New Database**" button.

The number [1] is referring to the button to browse for the initial database file. The number [2] button is enabling the user to upload the file to database of the program and start the running of this database file.

THE COMARK PROGRAM WINDOW

The main window of the application and contain the main functionalities of the program. the user can reach the window from “File”, then click on “Run CoMark” button. The window contains the list of the species that loaded from database file and the buttons to run the other program, construct the whole-proteome tree, and to view the final tree. Figure 3-4 is a screenshot and description of the window.

In the software, the user is capable of switching between windows using "**File**" button in the toolbar of the main program. The "**Help**" button in the toolbar of the main program will open the user manual (the file is PDF format).

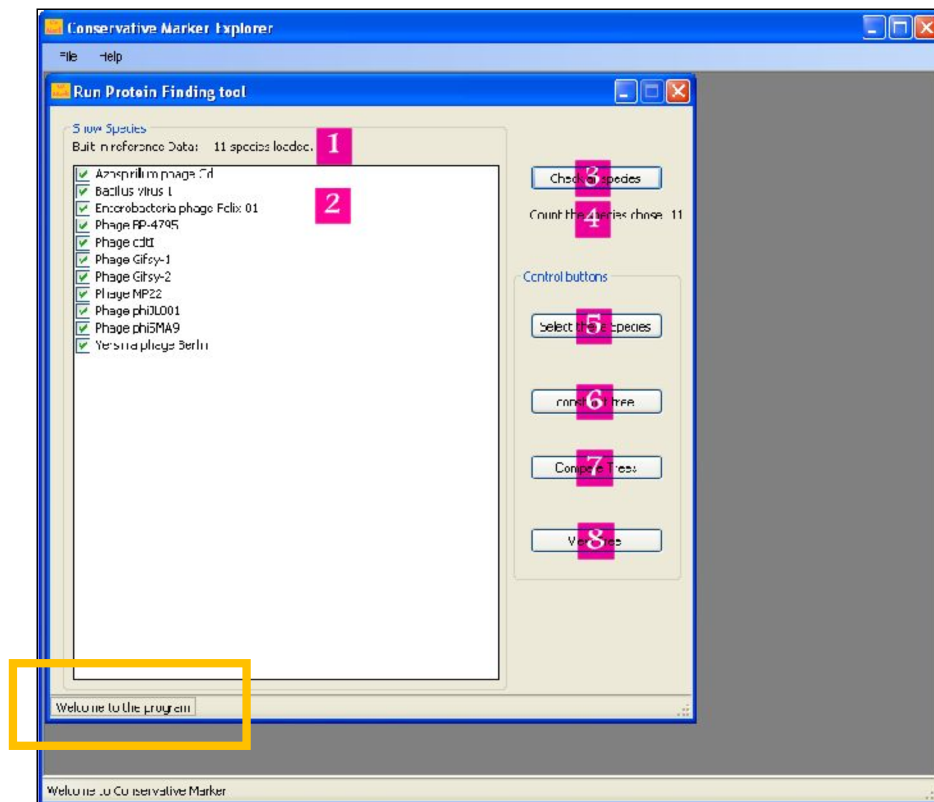


Figure 3-4. The screenshot shows the main interface of the program. The window is the main CoMark software window. The window contains the main functionality of the program. The user can get this window by clicking “**File”** in the toolbar, then, clicking “**Run CoMark**” button. Moreover, the status bar contains tips and guide of the program.

The number [1] is the number of species that loaded from the database file. The program initially loaded by sample database file which contains 11 species of phages. [2] In the program loading the species names will appear in a check list box. By this control box, the user prompts to choose the species of interest. [3] The user is capable of choosing all the species by clicking the "**Check all species**" button. [4] The number of species has

chosen and written in this line. [5] The "**Select the species**" button will start the writing a new database file of the selected species. [6] To start running the sequence aligning using **BlastP** and **ClustalW** programs, then to calculate the distance matrices, construct the individual trees of each dataset and finally construct the whole tree. [7] The button of "**Compare trees**" is to run **TOPD/FMTS** and compare the individual trees with the whole one. [8] The "**View Tree**" is to visualize the whole tree that is constructed.

The yellow box shows a line in the status bar. The line contains some instructions while running the program.

4. THE FINAL OUTPUT OF THE SOFTWARE

The application creates folder called "**CoMark**" in C partition on the hard disk. In case of run new instance of program and the folder is found, the program fires a warning message to notify that this directory will be deleted.

The final output is written in "C:\CoMark\" directory. The file "C:\CoMark\FinalTree.tre" is the whole-proteome tree of the species used in the study. The text file "PMResults.txt" in the same directory is the statistics of the nodal, split, and disagreement methods and the interrelation between the trees' distances under each method.

The other additional files are placed in a directory called "C:\CoMark\Extra_Files\". This directory included other files used by program. This directory created and left intentionally to help the user to check the output files and results if needed. The directory includes the following:

The result of each method is written in the text file. The resultant file name is related to each method name; for Disagreement the file name is "Final_Disagree_Results.txt", for nodal is "Final_Nodal_Results.txt", while split method output file called "Final_Split_Results.txt".

Each file is a table contains the names of the trees, the distance resulted from the methods used, and finally the mean of the pruned to unpruned tree.

The "SpID_with_Names.txt" file is a text file contains the number of strains in the input file and the name of each strain with the ID of each species to facilitate the use of it by the program during the run time.

The "TreesNamesFile.intopd" file can be opened by the text file editor and contains the TreeIDs and the distances of the tree.

The file named "TreesNamesFile_Explian.intopd" is also can be opened by the text file editor and contains the name of the trees of each dataset and the TreeID for each the main (whole-proteome) tree is the first one.

5. THE SIGNIFICANCE OF THE SOFTWARE

The application is studying the phylogeny and proteome of species with high rate of mutations. The software will be beneficial in some cases like:

1. This approach will help to identify a phylogenetic marker for some unclassified species and the species whose genome is a complicated one.

2. The study of phylogenetic markers, common, and conservative proteins is very crucial to the field of biology. These approaches will solve many problems in the metagenomic era and the tree of life.

3. An important trend in the phylogenetic studies is to track the history of the genes. Using the tree comparison method, the study is able to track the most common protein and it is identified as a most conservative protein.

4. The study succeeds to identify a common protein by using a simple database of proteome of certain species.

5. One of the most applications of the study is to refine the genes and eliminate the horizontally transferred genes and present the most conservative proteins.

6. The future improvements, like integrating additional functionalities, tree constructing techniques, detect/quantify HGT and/or detecting the gene order, are interesting research pipeline in the field of phylogenetic.

7. The software is user-friendly, and freeware with full online documentations.

APPENDIX

Appendix-1. Full list of species and strains of bacteriophages that being used in this study to construct the phage protein database.

Species Names	Strains' Names
EFAP-1 phage	EFAP-1 phage
Listeria phage	Listeria phage 2389
	Listeria phage A006
	Listeria phage A118
	Listeria phage A500
	Listeria phage B025
	Listeria phage B054
	Listeria phage P35
Bacillus phage	Bacillus phage BCJA1c
	Bacillus phage Cherry
	Bacillus phage Fah
	Bacillus phage Gamma
	Bacillus phage IEBH
	Bacillus phage phi105
	Bacillus phage SPBc2
	Bacillus phage SPP1
	Bacillus phage WBeta

Lactobacillus phage	Lactobacillus phage A2
	Lactobacillus phage KC5a
	Lactobacillus phage Lc-Nu
	Lactobacillus phage LL-H
	Lactobacillus phage Lrm1
	Lactobacillus phage phiadh
	Lactobacillus phage phiAT3
	Lactobacillus phage phig1e
	Lactobacillus prophage Lj928
	Lactobacillus prophage Lj965
Streptococcus phage	Streptococcus phage 2972
	Streptococcus phage 7201
	Streptococcus phage 858
	Streptococcus phage DT1
	Streptococcus phage MM1
	Streptococcus phage O1205
	Streptococcus phage phi3396
	Streptococcus phage Sfi11
	Streptococcus phage Sfi19
	Streptococcus phage Sfi21
	Streptococcus phage SM1
	Streptococcus thermophilus bacteriophage Sfi11

Lactococcus phage	Lactococcus phage 1706
	Lactococcus phage bIBB29
	Lactococcus phage bIL170
	Lactococcus phage bIL285
	Lactococcus phage bIL286
	Lactococcus phage bIL309
	Lactococcus phage bIL310
	Lactococcus phage bIL311
	Lactococcus phage bIL312
	Lactococcus phage BK5-T
	Lactococcus phage c2
	Lactococcus phage jj50
	Lactococcus phage P008
	Lactococcus phage P335 sensu lato
	Lactococcus phage phiLC3
	Lactococcus phage Q54
	Lactococcus phage r1t
	Lactococcus phage sk1
	Lactococcus phage TP901-1
	Lactococcus phage Tuc2009
	Lactococcus phage ul36
Staphylococcus phage	Staphylococcus aureus phage phiNM3
	Staphylococcus phage 11
	Staphylococcus phage 187
	Staphylococcus phage 2638A
	Staphylococcus phage 29
	Staphylococcus phage 37

Staphylococcus phage 3A
Staphylococcus phage 42E
Staphylococcus phage 47
Staphylococcus phage 52A
Staphylococcus phage 53
Staphylococcus phage 55
Staphylococcus phage 69
Staphylococcus phage 71
Staphylococcus phage 77
Staphylococcus phage 80alpha
Staphylococcus phage 85
Staphylococcus phage 88
Staphylococcus phage 92
Staphylococcus phage 96
Staphylococcus phage CNPH82
Staphylococcus phage EW
Staphylococcus phage PH15
Staphylococcus phage phi 12
Staphylococcus phage phi13
Staphylococcus phage phiETA
Staphylococcus phage phiETA2
Staphylococcus phage phiETA3
Staphylococcus phage phiMR11
Staphylococcus phage phiMR25
Staphylococcus phage phiN315
Staphylococcus phage phiNM
Staphylococcus phage phiPVL108
Staphylococcus phage phiSLT

	Staphylococcus phage PVL
	Staphylococcus phage X2
	Staphylococcus prophage phiPV83

Appendix-2. List of proteins belong to three different of the datasets. These dataset contains largest number of proteins compared to other triple occurred datasets (those trees have the lowest 1% distance scores with the three methods at the same time). The “X” sign means this protein is absent in this dataset. In general, most of the proteins are similar between the resultant 3 datasets.

Dataset of protein [gi|orf8 protein from EFAP-1 phage]	**Dataset of protein [putative tape measure protein from Lactobacillus phage Lc-Nu] gi|78000025|ref|YP_358771.1**	**Dataset of protein [phage tail tape measure protein from Staphylococcus aureus phage phiNM3] gi|118725105|ref|YP_908841.1**
77ORF001	77ORF001	77ORF001
carbamoyl-phosphate synthase large subunit	carbamoyl-phosphate synthase large subunit	carbamoyl-phosphate synthase large subunit
gp14	gp14	gp14
gp15	X	gp15
X	hypothetical protein Ljo_1425	X
hypothetical protein Ljo_1425d	X	hypothetical protein Ljo_1425d
hypothetical protein LPLLH_ORF360	hypothetical protein LPLLH_ORF360	hypothetical protein LPLLH_ORF360

hypothetical protein phiadhp52	hypothetical protein phiadhp52	X
hypothetical protein phiETA3_gp55	hypothetical protein phiETA3_gp55	hypothetical protein phiETA3_gp55
hypothetical protein phiSLTp50	hypothetical protein phiSLTp50	hypothetical protein phiSLTp50
hypothetical protein PVL_15	X	hypothetical protein PVL_15
X	hypothetical protein PVL_16	X
hypothetical protein SA1766	hypothetical protein SA1766	hypothetical protein SA1766
hypothetical protein sk1p14	hypothetical protein sk1p14	hypothetical protein sk1p14
hypothetical protein Tuc2009_46	hypothetical protein Tuc2009_46	hypothetical protein Tuc2009_46
minor capsid protein	minor capsid protein	minor capsid protein
minor tail protein	minor tail protein	minor tail protein
ORF001	ORF001	ORF001
ORF013	ORF013	X
orf19	orf19	orf19

ORF33	ORF33	ORF33
PblA	PblA	PblA
phage tail tape measure protein	phage tail tape measure protein	phage tail tape measure protein
X	phage tape measure protein	phage tape measure protein
phi PVL ORF 15 and 16 homologue	X	phi PVL ORF 15 and 16 homologue
X	phi PVL ORF 17 homologue	X
putative minor tail protein	putative minor tail protein	putative minor tail protein
X	putative phage tail tape measure protein	putative phage tail tape measure protein
putative phage tape measure protein	putative phage tape measure protein	putative phage tape measure protein
putative phage-related tail tape measure protein	putative phage-related tail tape measure protein	putative phage-related tail tape measure protein
putative putative minor tail protein	putative putative minor tail protein	putative putative minor tail protein

putative tail component protein	putative tail component protein	putative tail component protein
putative tail lysin	putative tail lysin	putative tail lysin
putative tail protein	putative tail protein	putative tail protein
putative tail tape measure protein	putative tail tape measure protein	putative tail tape measure protein
putative tape measure protein	putative tape measure protein	putative tape measure protein
putative tape-measure protein	putative tape-measure protein	putative tape-measure protein
X	putative transglycosylase	putative transglycosylase
Q-orf8 [EFAP-1 phage]	Q-orf8 [EFAP-1 phage]	Q-orf8 [EFAP-1 phage]
similar to phage bIL170 l16	similar to phage bIL170 l16	similar to phage bIL170 l16
structural protein	structural protein	structural protein
X	tail adsorption protein	X
tail fiber protein	tail fiber protein	tail fiber protein
tail length tape measure protein	tail length tape measure protein	tail length tape measure protein

tail length tape measure protein2	tail length tape measure protein2	tail length tape measure protein2
tail protein	tail protein	tail protein
tail tape measure	tail tape measure	tail tape measure
tail tape measure protein	tail tape measure protein	tail tape measure protein
tape measure protein	tape measure protein	tape measure protein
Tmp	Tmp	TMP
unknown	unknown	unknown

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool, *Journal of molecular biology*, **215**, 403-410.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389.

Andersson, J.O. (2005) Lateral gene transfer in eukaryotes, *Cellular and Molecular Life Sciences*, **62**, 1182-1197.

Bacteriodetes, C. and Spirochaetes, F. Signatures for Archaea Groups.

Brown, J. (2003) Ancient horizontal gene transfer, *Nature Reviews Genetics*, **4**, 121-132.

Creevey, C., Fitzpatrick, D., Philip, G., Kinsella, R., O'Connell, M., Pentony, M., Travers, S., Wilkinson, M. and McInerney, J. (2004) Does a tree–like phylogeny only exist at the tips in the prokaryotes?, *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **271**, 2551.

Davids, W. and Zhang, Z. (2008) The impact of horizontal gene transfer in shaping operons and protein interaction networks – direct evidence of preferential attachment, *BMC Evolutionary Biology*, **8**, 23.

Doolittle, W. (1999) Phylogenetic classification and the universal tree, *Science*, **284**, 2124.

Eisen, J. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis, *Current Opinion in Genetics & Development*, **10**, 606-611.

Estabrook, G., McMorris, F. and Meacham, C. (1985) Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units, *Systematic Biology*, **34**, 193.

Fall, S., Mercier, A., Bertolla, F., Calteau, A., Gueguen, L., Perriï¨re, G., Vogel, T.M. and Simonet, P. (2007) Horizontal Gene Transfer Regulation in Bacteria as a â€œSpandrelâ€ of DNA Repair Mechanisms, *PLoS ONE*, **2**, e1055.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, **39**, 783-791.

Felsenstein, J. (1990) PHYLIP package, v3. 3, *Department of Genetics, University of Washington, Seattle*.

Filée, J., Tétart, F., Suttle, C. and Krisch, H. (2005) Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the

biosphere, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 12471.

Fouts, D. (2006) Phage_ Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences, *Nucleic Acids Research*, **34**, 5839.

Fox, G., Stackebrandt, E., Hespell, R., Gibson, J., Maniloff, J., Dyer, T., Wolfe, R., Balch, W., Tanner, R. and Magrum, L. (1980) The phylogeny of prokaryotes, *Science (New York, NY)*, **209**, 457.

Fuhrman, J. (1999) Marine viruses and their biogeochemical and ecological effects, *Nature*, **399**, 541-548.

Fuller, N.J., Wilson, W.H., Joint, I.R. and Mann, N.H. (1998) Occurrence of a Sequence in Marine Cyanophages Similar to That of T4 g20 and Its Application to PCR-Based Detection and Quantification Techniques, *Appl. Environ. Microbiol.*, **64**, 2051-2060.

Garcia-Vallve, S., Guzman, E., Montero, M. and Romeu, A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes, *Nucleic Acids Research*, **31**, 187.

Gupta, R. (2005) Molecular sequences and the early history of life, *Microbial Phylogeny and Evolution: Concepts and Controversies*, 160-183.

Hambly, E., Tetart, F., Desplats, C., Wilson, W., Krisch, H. and Mann, N. (2001) A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2, *Proceedings of the National Academy of Sciences*, **98**, 11411.

Henz, S., Huson, D., Auch, A., Nieselt-Struwe, K. and Schuster, S. (2005) Whole-genome prokaryotic phylogeny, *Bioinformatics*, **21**, 2329.

Jiang, S. and Paul, J. (1998) Gene transfer by transduction in the marine environment, *Applied and environmental microbiology*, **64**, 2780.

Kidambi, S., Ripp, S. and Miller, R. (1994) Evidence for phage-mediated gene transfer among Pseudomonas aeruginosa strains on the phylloplane, *Applied and Environmental Microbiology*, **60**, 496.

Koonin, E., Makarova, K. and Aravind, L. (2001) H ORIZONTAL G ENE T RANSFER IN P ROKARYOTES: Quantification and Classification 1, *Annual Reviews in Microbiology*, **55**, 709-742.

Krauss, V., Thummler, C., Georgi, F., Lehmann, J., Stadler, P.F. and Eisenhardt, C. (2008) Near Intron Positions Are Reliable Phylogenetic Markers: An Application to Holometabolous Insects, *Mol Biol Evol*, **25**, 821-830.

Kubicka, E., Kubicki, G. and McMorris, F.R. (1995) An algorithm to find agreement subtrees, *Journal of Classification*, **12**, 91-99.

Kurland, C., Canback, B. and Berg, O. (2003) Horizontal gene transfer: a critical view, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9658.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E. and Raoult, D. (2008) The virophage as a unique parasite of the giant mimivirus, *Nature*, **455**, 100-104.

Lane, D., Pace, B., Olsen, G., Stahl, D., Sogin, M. and Pace, N. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses, *Proc. Natl. Acad. Sci. USA*, **82**, 6955-6659.

Lawrence, J., Hatfull, G. and Hendrix, R. (2002) Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches, *Journal of bacteriology*, **184**, 4891.

Le Marrec, C., Van Sinderen, D., Walsh, L., Stanley, E., Vlegels, E., Moineau, S., Heinze, P., Fitzgerald, G. and Fayard, B. (1997) Two groups of bacteriophages infecting Streptococcus thermophilus can be distinguished on the basis of mode of packaging and genetic determinants for major structural proteins, *Applied and environmental microbiology*, **63**, 3246.

Liao, W., Song, S., Sun, F., Jia, Y., Zeng, W. and Pang, Y. (2008) Isolation, characterization and genome sequencing of phage MZTP02 from Bacillus thuringiensis MZ1, *Archives of Virology*, **153**, 1855-1865.

Lorenz, M. and Wackernagel, W. (1994) Bacterial gene transfer by natural genetic transformation in the environment, *Microbiology and Molecular Biology Reviews*, **58**, 563.

Mirkin, B., Fenner, T., Galperin, M. and Koonin, E. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes, *BMC Evolutionary Biology*, **3**, 2.

Nelson, D. (2004) Phage taxonomy: we agree to disagree, *Journal of bacteriology*, **186**, 7029.

Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation, *Nature*, **405**, 299-304.

Pace, N., Stahl, D., Lane, D. and Olsen, G. (1986) The analysis of natural microbial populations by ribosomal RNA sequences, *Advances in microbial ecology*, **9**, 1-55.

Pedersen, M., Østergaard, S., Bresciani, J. and Vogensen, F. (2000) Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901-1 involved in tail length determination and baseplate assembly, *Virology*, **276**, 315-328.

Penny, D. and Hendy, M.D. (1985) The Use of Tree Comparison Metrics, *Systematic Zoology*, **34**, 75-82.

Philippe, H. and Douady, C. (2003) Horizontal gene transfer and phylogenetics, *Current Opinion in Microbiology*, **6**, 498-505.

Proux, C., Van Sinderen, D., Suarez, J., Garcia, P., Ladero, V., Fitzgerald, G., Desiere, F. and Brussow, H. (2002) The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria, *Journal of bacteriology*, **184**, 6026.

Puigbo, P., Garcia-Vallve, S. and McInerney, J. (2007) TOPD/FMTS: a new software to compare phylogenetic trees, *Bioinformatics*, **23**, 1556.

Ragan, M. (2001) Detection of lateral gene transfer among microbial genomes, *Current Opinion in Genetics & Development*, **11**, 620-626.

Robinson, D.F. and Foulds, L.R. (1981) Comparison of phylogenetic trees, *Mathematical Biosciences*, **53**, 131-147.

Rohwer, F. and Edwards, R. (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage, *Journal of bacteriology*, **184**, 4529.

Smalla, K., Borin, S., Heuer, H., Gebhard, F., van Elsas, J. and Nielsen, K. (2000) Horizontal transfer of antibiotic resistance genes from transgenic plants to bacteria, *Canadian Wheat Board Agrium Foragen*, 146.

Snel, B., Bork, P. and Huynen, M. (1999) Genome phylogeny based on gene content, *Nature genetics*, **21**, 108-110.

Steel, M.A. and Penny, D. (1993) Distributions of Tree Comparison Metrics-Some New Results, *Systematic Biology*, **42**, 126-141.

SUSSKIND, M. and BOTSTEIN, D. (1978) Molecular Genetics of Bacteriophage P22t, *Microbiology and Molecular Biology Reviews*, 385-413.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0, *Molecular biology and evolution*, **24**, 1596.

Tetart, F., Desplats, C., Kutateladze, M., Monod, C., Ackermann, H.-W. and Krisch, H.M. (2001) Phylogeny of the Major Head and Tail Genes of the Wide-Ranging T4-Type Bacteriophages, *J. Bacteriol.*, **183**, 358-366.

Thompson, F., Gevers, D., Thompson, C., Dawyndt, P., Naser, S., Hoste, B., Munn, C. and Swings, J. (2005) Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis, *Applied and environmental microbiology*, **71**, 5107.

Thompson, J. and Higgins, D. (1994) Gibson TJ CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, **22**, 4673-4680.

Twort, F. (1915) AN INVESTIGATION ON THE NATURE OF ULTRA-MICROSCOPIC VIRUSES, *The Lancet*, **186**, 1241-1243.

Wolf, Y., Rogozin, I., Grishin, N., Tatusov, R. and Koonin, E. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades, *BMC Evol Biol*, **1**, 1471-2148.

Zeigler, D. (2003) Gene sequences useful for predicting relatedness of whole genomes in bacteria, *International Journal of Systematic and Evolutionary Microbiology*, **53**, 1893.

SUMMARY IN KOREAN

생물종에 있어서 계통학적 마커를 찾는 것은 종간의 진화적 관계를 밝히는 데 매우 중요한 역할을 한다. 전장 유전체 비교에 대한 많은 방법이 있으며, 각 비교 방법에 따른 제약이 존재한다. 뿐만 아니라 박테리오파지의 유전체는 이러한 방법들에 의해 분석하는데 어려움을 갖는다. 본 연구는 계통학적 마커의 자동적이고 더 적합한 예측 시스템을 제공하는 것을 목표로 한다.

본 연구를 통하여 전장 유전체에 걸친 보존적인(conservative) 단백질을 이론적으로 예측하는 "CoMark"라는 새로운 응용 프로그램을 개발하였다. 이러한 방법을 통해 밝혀진 단백질 서열들은 해당 종의 계통학적 마커로서 적용될 수 있다.

예측 시스템은 해당 종의 전체 단백질 서열을 서열 유사도에 근거하여 많은 그룹으로 단백질 서열을 나누는 것으로부터 시작된다. 각 그룹은 배열(alignment)을 통해 그룹별 단백질간의 유사도를 계산하고, 최종적으로 종간의 평균적인 유사도를 계산하여 종간의 거리는 결정한다. 이렇게 생성된 유전체 계통수는 각 그룹에서 계산된 모든 계통수와 비교하는 과정을 거치게 된다. 실험 자료로서 하나의 쿼리종을 포함한 97 개의 종에 대한 5821 개의 단백질 서열이 이용되었다.

본 추정시스템은 유전적 마커를 추정하기 위해 단백질체적 관점을 적용하였으며, 전장 유전체 비교에 근거한 마커를 탐지할 수 있었다. 비록 본 연구는 박테리오파지에 초점을 맞추었지만, 이러한

추정방법은 파지와 유사한 특정을 지닌 미생물 유전체에 적용될 수 있을 것이다.

개발된 소프트웨어는 무료이며 Windows 용 데스크톱 시스템에서 사용할 수 있도록 개발되었다. 관련 자료와 사용자 안내는 http://snugenome.snu.ac.kr/comark/ 을 통하여 제공된다.

Keywords: Bacteriophage, conservative proteins, horizontal gene transfer, phylogenetic marker, proteins, tree comparison, whole-proteome tree.

학번: 2008-22515